ENDEAVOUR

A quarterly review, published in five languages,
designed to record the progress of the sciences
in the service of mankind

VOLUME XX July 1961 NUMBER 79

CONTENTS

5.001.891:070.432
Editorial: Investing in Research 119

530.12:531.5
Gravitational Waves — H. BONDI, M.A., F.R.S. — 121

551.593.51
The Green Flash and Kindred Phenomena — D. J. K. O'CONNELL, S.J., D.Sc. — 131

541.653
Optical Rotatory Dispersion — CARL DJERASSI, Ph.D., D.Sc. — 138

576.312:577.15
The Nucleolus of the Cell Nucleus — J. L. SIRLIN, Dr.Nat.Sc. — 146

668.81:547.458.8
Reactive Dyestuffs for Cellulose — I. D. RATTEE, B.Sc., A.R.C.S. — 154

623.454.7(091):93
Greek Fire — J. R. PARTINGTON, M.B.E. (Military), D.Sc. — 162

049.3
Book Reviews 166

Short Notices of Books 171

Notes on Contributors 172

Some Books Received .. .. Cover iii

Editor: TREVOR I. WILLIAMS, M.A., B.Sc., D.Phil., F.R.I.C.
Assistant Editor: G. N. J. BECK, B.Sc., A.M.I.E.E.
Editorial Assistant: W. A. OSMAN, B.Sc., M.Sc., A.R.C.S.

Imperial Chemical Industries Limited, Millbank, London, S.W.1
Communications to the Editor should be addressed to North Block, Thames House, Millbank, London, S.W.1

RECEIVED
26 JUN 1961
LIBRARY
WIDNES LABORATORY
WIDNES

Investing in research

The pursuit of knowledge for its own sake has long been represented as the highest form of human endeavour, and this is certainly a proposition with which few scientists will quarrel. Moreover, it can be argued that scientific instinct and the public interest go hand in hand, for there is no lack of examples of discoveries made without thought of application that have in due course proved of the greatest practical importance. On the evidence of past experience it is undeniable that, in the long run, scientific research is bound to produce useful results, even though their nature may be unpredictable. This is a comforting situation, for it may be held to be sufficient justification for almost any sort of inquiry. Nevertheless, the total expenditure on research is now so great that those concerned with its allocation must increasingly consider the quantitative as well as the qualitative side of the argument. Exact estimates of research expenditure are impossible to make, because one can neither define research precisely nor follow its ultimate ramifications. Approximate calculations are feasible, however, and a recent estimate by the Advisory Council on Scientific Policy shows that British expenditure in 1959 was £478 million, representing 2·35 per cent of the gross national product. Of this total, rather more than half represented research and development in private industry and almost exactly one-third was government expenditure, mainly on defence; university research represented rather less than five per cent of the total. Although the United States total, calculated on the same basis, is many times greater (£3971 million), the percentage of the gross national product is much the same, namely 2·74 per cent. Strict comparisons with other countries are not possible, but there is good reason to suppose that expenditure in Russia is relatively at least as high, if not higher, and most of the other leading nations also devote very large sums to scientific research. Nevertheless, it is still generally acknowledged that still more must be spent if we are to make the best use of scientific possibilities. How great an advantage the further vast expenditure contemplated will be depends to a great extent upon its allocation: if well spent, the greater total may, by virtue of results, represent a smaller percentage of the gross product.

The present world expenditure on research and development is enormous, both absolutely and if measured against the comparable figures for the years before the first world war, when so many still-current ideas about policy were formed. At that time science was generally far more remote from everyday life than it is today, and if the research worker had to justify himself at all it was only to those who supported him. Today, however, the situation is very different, for much of the money comes directly or indirectly from public sources. In the long run it is essential, if support is to be continued and increased, to convince the policy-forming section of the public that it is receiving value for money. Rightly or wrongly, in this respect value for money means sufficient practical application—considered in the widest sense of a higher standard of living, better health, defence, national prestige, and so on—to justify the overall outlay. A substantial part of research must now be financed from the wealth that its application creates, for the total sum involved has become far too great to be spent without thought of practical advantage. This is the hard reality of the situation, unpalatable though it may be to those who believe that scientists should be the only arbiters of things scientific. If science is not merely to flourish on its present scale but to expand, public support and understanding of its aims are absolutely essential.

This does not, of course, at all imply that we must do more and more applied research to the detriment of fundamental work, for, apart from considerations of scholarship, that would ultimately be ruinous. But it does mean that, as a matter of common prudence, more thought must be given to the factors that determine where the balance between these two fields of research should properly be struck. That there is a point of balance is evident. There are types of investigation that, so far as can be seen at the time, promise no useful applications whatever. In a balanced community such pursuits deserve support for the same sort of reasons as lead to encouragement of music, drama, literature, art, and other aesthetic interests; they are part of the general cultural background. At the other end of the scale, even the most convinced materialist can see that a term is set to the useful application of knowledge if we do not ensure that our store of it is constantly increased. But the public purse is not inexhaustible, and this in itself means that there can be only limited support for all kinds of research. Where should the balance lie?

Certainly no hard and fast answer can be given to this question, for, apart from other considerations, conditions vary greatly between the different branches of science. In some, progress is delayed for lack of fundamental knowledge; in others, knowledge is accumulating faster than it can be turned to practical account, and we have, as it were, capital earning no interest. Today, the different branches of science increasingly cross-fertilize each other, and it might seem to be logical to increase expenditure in the weakest fields so that progress may take place fairly evenly on a broad front. But this, if it were feasible at all, could be done only on a long-term basis, for merely making money available is only the beginning. Not only must laboratories and equipment be provided, but, far more difficult, the trained men must be found to use them. In science, as in other pursuits, economic pressures are heavy, and incentives can be given to encourage workers to enter fields where they are particularly needed. But this, again, is by no means the whole answer; what really produces results, in science as in so many other branches of co-operative endeavour, is enthusiastic leaders, and of these the supply is very limited.

But neither leaders nor led create themselves, and the education of the potential scientist is of the first importance in maintaining and increasing the number of active workers. On a long-term view this is something that may properly be considered as part of the cost of research, though it is too rarely looked upon as such; nevertheless, it is indisputable that the quality of a scientist owes much to his early training. At the moment, virtually every country in the world is making great and costly efforts to multiply its output of science graduates; yet illogically small attention is devoted to ensuring adequate teaching at school level, during the impressionable years in which the principles of the scientific method are most firmly inculcated. In most countries, little has been done to make the reward of those concerned with preparing the next generation of scientists comparable with that of working with the present one. Here, too, there is clearly a lack of balance, though it is easier to diagnose the fault than to suggest a remedy. Nevertheless, the mere realization that we are letting the edifice grow without paying sufficient attention to its foundations is an important step towards a reappraisal of the apportioning of scientific expenditure.

The need to strike a proper balance between pure and applied research has already been touched on, and it presents a problem of the greatest importance for the future of science. The belief dies hard that the former is in some way intellectually superior to the latter; during their training the best men tend to be imbued with the idea that the only worth-while goal is to remain in academic research, industry being represented as the refuge of the second-rate. Though the belief today is held only by a minority, this is sufficient to add to the difficulty of recruiting good men for applied research. This position would be justifiable if the basic premise were correct, for it would be wrong to use the best brains at anything less than their full capacity. Yet all the evidence is to the contrary; those who have experience of both have no doubt that for the very great majority of scientists research in industry can offer problems no less intellectually challenging than are encountered in academic laboratories. This does not mean, of course, that all industrial science is exciting, for necessarily much of it is of a routine nature. But much the same may be said of a good deal of academic research. The collecting of essential information may well be mundane and repetitive; it is in devising means of gaining new knowledge, and in the interpreting of observations, that the original mind can best exercise itself. In considering the general question of expenditure on research it is necessary to think not only how the total should be apportioned between pure and applied work but what steps need be taken to ensure that both get a proper share of the best workers. Until this is done, the full advantages of the application of new knowledge will be difficult to realize. Moreover, if the present prejudice lingers, the purists themselves will suffer, for it is undeniable that the wealth created by applied science makes a considerable contribution to sustaining pure science; in the long run, the fortunes of both are inextricably linked.

Gravitational waves

H. BONDI

The experimental study of gravitational waves poses great problems, chiefly because of the extreme weakness of gravitational fields; the waves could become appreciable only in the case of very massive bodies moving with high accelerations. Although there is no present prospect of our being able to observe gravitational waves, some of their principal properties can be inferred. This article shows how gravitational theory is related to the rest of physics, and reviews some of the lines along which it has developed in recent years.

Gravitational waves are not only unfamiliar, even by name, but are distinctly unlikely to be observed. Despite this, their study has received appreciable attention, and is instructive because it is linked to so many branches of science. It stands between the physics of large-scale phenomena and quantum theory, between gravitational theory and the subject of electromagnetic and other waves, between pure mathematics and physics. Present investigations are concerned with the gravitational fields of rapidly moving bodies. Gravitation is a negligible force unless the body exerting it is very massive, and high accelerations are virtually unknown among such bodies: hence the experimental intractability of the subject. One is necessarily confined to mathematical investigations, based on the best available theory of gravitation, namely Einstein's general theory of relativity; many of the difficulties of the subject are due to the mathematical complexities of this theory. To make clear why enough confidence is felt in this theory to justify the effort spent on gravitational waves, it is necessary to go far back into physics and consider the basic character of the theory. Because our subject is interwoven with so many other branches of physics, it will be necessary to review certain aspects of these in some detail before the problems of gravitational waves can properly be formulated.

NEWTONIAN DYNAMICS

Though it is well known that Newton's three laws put dynamics on a firm footing and enabled it to enjoy a period of rapid growth, the exact nature of his contribution is not always clearly understood. It will be recalled that Newton related force to acceleration, whereas previously it was believed to be related to velocity. Why was it such an important contribution to make what was effectively a new definition of force, in which acceleration replaced velocity? The reason is clearly seen when the motion of the Earth is considered. If one considers the force responsible for this motion and presumes it to be related to its velocity, as in pre-Newtonian days, then one should be looking in the direction of the Earth's velocity to find something responsible for its motion. But looking in the direction of the Earth's velocity is fruitless; sometimes one sees one rather insignificant-looking star, then another, and at other times virtually nothing. If, however, one looks in the direction of the Earth's acceleration one sees the Sun, and it is very acceptable indeed to regard this most important body as being in some way the cause of the orbit of the Earth. To put it a little differently, Newton's definition of force enables one to link a kinematical observable, the acceleration of the Earth, with something that can be observed optically, the direction of the Sun. It is a characteristic of any successful scientific hypothesis that it links two previously unrelated kinds of observation, and this is achieved here; it is in this way that Newton's dynamics proved so eminently successful. Moreover, we observe here a point that will recur several times in this article, namely the essential unity of physics. Physics cannot be subdivided into a number of watertight compartments, such as dynamics, optics, and heat; these fields are all linked with each other in numerous ways, and many of the most important physical insights would be lost if the links were not borne in mind. We cannot have one type of physics for dynamics and another type of physics for another subject; dynamics could never have been established without optical observations, and there is no such thing as a purely dynamical experiment. More than one branch of physics is always concerned, and, as we have just noted, the very foundation of dynamics depends on the union of a dynamical and an optical observation.

The next point to which attention must be directed is the nature of the system of reference relative to which acceleration is measured. Force is regarded in Newtonian dynamics as something essentially real, something of true physical significance. Newton's second law implies, therefore, that acceleration is similarly real. But acceleration is the rate of change of velocity, and velocity is the rate of change of position, and position, velocity, and acceleration are all in some way relative. The state of motion of the measuring system decides what values one obtains for each of these variables. Thus, if acceleration is something physically real, something that can be identified with a true physical variable, force, then only those systems of reference that give acceleration the correct value are admissible for our observations. If we have such a system of reference and take a second system of reference in unaccelerated motion relative to it—that is, moving with constant velocity in a fixed direction—then the velocities measured in the second system will differ from the velocities measured in the first system, but the accelerations measured in the second system will be the same as those measured in the first system. The second system will therefore be just as admissible as a standard reference system as the first one. Therefore, in Newtonian dynamics we do not have a unique system of reference, we do not have some kind of standard of rest for the Universe at large, but we have a whole family of such systems of reference, any two members being in relative unaccelerated motion. Any such system of reference is called an inertial system, and acceleration as observed from an inertial system is connected with force through Newton's second law. There is therefore a very sharp distinction in Newtonian dynamics between two kinds of frames of reference, namely inertial frames and non-inertial; only the forces measured in inertial frames are regarded as real. If we work in an accelerated frame of reference, that is, a non-inertial one, then fictitious forces arise. The best known of these is the centrifugal force, which arises when we work in a rotating system of coordinates, and there are several others of that kind.

Since all forces are found to be the same, from whichever inertial system they are viewed, it follows that all inertial systems are dynamically completely equivalent. One can therefore state the Newtonian principle of relativity in the following way:

1. There exists a sharp distinction between inertial systems and non-inertial systems.
2. Any two inertial systems are in relative unaccelerated motion.
3. Any two inertial systems are dynamically completely equivalent, and it is impossible by any criterion of internal dynamical experiments to distinguish one from another.

The significance of this last principle is simply that if, for example, we are in a smoothly moving aircraft and conduct experiments within this aircraft, then all these experiments will give precisely the same result as such experiments would give in a room on the ground. As there are no physical effects of velocity, but only effects of acceleration, it follows that one cannot discover the velocity of any object by internal experiments. Velocity is not absolute but purely relative, whereas acceleration is absolute.

In addition to these basic principles, simple mathematical rules of transformation can be set up that relate measurements made in one system to measurements made in another, whenever this is required; that is, when objects external to both these systems are being measured. In these transformations, acceleration is naturally preserved, and also time, which in Newtonian physics is considered as a universal and absolute quantity.

ELECTROMAGNETISM AND SPECIAL RELATIVITY

In Newtonian dynamics, velocity is never regarded as something fundamental; indeed, it is, in a certain sense, arbitrary. In any field of force, in any conceptual physical situation, we can start off at a given moment by assigning arbitrary positions and velocities to all the particles. The laws of dynamics then specify their subsequent behaviour. In other branches of physics, however, and especially in electromagnetism and the associated field of optics, there exists a velocity of fundamental importance, namely the velocity of light. Not only is this the velocity in a vacuum of all the many kinds of electromagnetic waves, but it also occurs in the basic equations of electromagnetic induction. If the spirit of the principles of Newtonian relativity is applied, it follows that this basic feature, the velocity of light, should be independent of the inertial system from which it is observed. The assertion that this is so, and, therefore, the re-affirmation of the Newtonian principles of relativity for the electromagnetic case, is known as the special theory of relativity. Stress is so often laid on the revolutionary character of this theory that it is frequently overlooked that its re-affirmation that Newtonian ideas are just as applicable to electromagnetism as to dynamics is simply an assertion of the validity of Newtonian relativity for the whole of physics. Thus the theory is, in a sense, profoundly conservative; the principles that have been found to hold for dynamics are stated to hold for the whole of physics. The special theory of relativity is a necessary consequence of any assertion that the unity of physics is essential, for it would be intolerable for all inertial systems to be equivalent from a dynamical point of view yet distinguishable by optical measurements. It now seems almost incredible that the possibility of such a discrimination was taken for granted in the nineteenth century, but at the time it was not easy to see which was more important—the universal validity of the Newtonian principle of relativity or the absolute nature of time, referred to previously.

A simple piece of mathematics shows that the velocity of light cannot be the same in all inertial frames of reference if the assumption of a universal time is maintained. One has to drop the metaphysical idea that time is something absolute, and regard time in the proper physical sense as that which is measured by a clock. In modern science it is becoming more and more understood that it is pointless to talk about quantities unless it is stated how they are to be measured. It is not scientific to regard uniform even-flowing time as being in some mysterious way unalterable. Nor can we assume that, in some ill-defined sense, the time measured by a clock in one system of reference equals the time measured in another system. What we can assert by taking over the Newtonian principle of relativity, which in its application to the whole of physics is called Einstein's special principle of relativity, is the equivalence of all inertial systems. New transformation rules have to be stated, and their essential quality is that time becomes linked with space. This is generally referred to as the four-dimensional nature of space-time. Although this phrase is sometimes regarded as mysterious, the basic idea is perfectly straightforward. What we mean by the three-dimensional nature of space is that we require three different measurements to specify the extension of a body, but that by a rotation they may become mixed up with each other. They are not in any sense absolute, although we may work out from them quantities, such as the distance between two points within the body, that are independent of the way in which we look at it. We mean by the four-dimensional nature of space-time that the velocity of the observer may lead to a mixing up of time-interval with the space-extent of an object in just this manner. The essential point is that this effect is very similar to that of a rotation on the three space-coordinates, and that just as certain measurements, such as the distance between two points of the body, are independent of the aspect, so certain other types of measurement, referred to as the interval, are independent of the state of motion of the observer.

It is not necessary here to refer to the innumerable experimental tests of the theory of special relativity or to its many well-known achievements, such as the identification of mass and energy. It will, however, be of advantage to consider its algebraic power, which arises directly from the four-dimensional nature of space-time. There are few simple algebraic structures that fit into space-time, and this fact enables very important deductions to be made. We recognize that in ordinary space of three dimensions the possibility of a change of position of the observer reduces the number of physical laws that are permissible. It is permissible to say that the force on a moving charged particle due to a magnetic field is at right angles to the velocity of the particle and to the direction of the magnetic field. But to say that the force is always in a direction to the right of the particle is not permissible, because this would mean different things to different observers.

Let us now imagine a stage in the progress of science when the laws of propagation of light were well known, enabling the special theory of relativity to be established, but when the laws of electromagnetism were still very little known and the electromagnetic nature of light was not understood at all. Suppose, in particular, that magnetism were wholly unknown and that only the following two simple laws of electrostatics had been discovered. Firstly, that the lines of electric force originate only at electric charges and that this is true whether the charges are moving or not. Secondly, that the electric field is conservative—meaning that if you take a charge round a closed circuit the total work you can get out of the electric field vanishes—but only if the field is constant in time. How would the situation look from a relativistic point of view? In relativity, we have a space of four dimensions, and therefore every directed quantity (vector) has to have four components that can, to some extent, be transformed into each other by changing the velocity of the observer. Thus in dynamics the momenta are the three space-components of such an entity and energy is the time-component. Throughout the application of relativity one usually finds that the relevant time-component is associated with the three space-components of a vector. However, the law concerning charges and the origin of electric lines of force is not of this kind. Not only does it involve just the three space-components of the electric field, but its validity, even when charges are moving, explicitly denies the existence of a time-component. It follows from this that the electric field cannot be such a simple algebraic structure as the momentum and energy of a particle constitute. Something more complicated has therefore to be examined, and the next in order of complication after a vector is a tensor.

A tensor is an entity having, in four dimensions, sixteen components; however, the number of these components may be reduced by symmetry considerations. In a particularly simple form of symmetry (antisymmetry) the number of components is reduced to six, three of which form one group and three another, each being transformed into the other by motion. The three components of the first group can readily be identified with the three space-components of the electric field and mathematically cannot have a time-component associated with them. The three relevant components of an antisymmetric tensor are the only simply occurring set with this property. Associated with them are the three other components, and as these arise from velocity, our relativist could deduce that, just as electric charges at rest produce an electric field, electric charges in motion (an electric current) should produce another kind of field. This is what we call the magnetic field, and the relativist would have deduced its existence from simple considerations of symmetry. Moreover, the fact that the electric field in its static condition was conservative will have enabled him to go a step further in identifying the structure of the electromagnetic field, for he would have been able to deduce that the entire field can be derived from a potential vector in four dimensions. Then, linking the field, the entire complexity of which he could thus have deduced from relativity considerations, with the sources of the field (the charges at rest and in motion), he would have arrived at the full set of Maxwell's equations describing the electromagnetic field. He could then have deduced from these equations that changes in the electromagnetic field would be radiated and propagated with the velocity of light. To complete his set of equations he would naturally have to make an assumption about causality, that is, that news of changes in a system travels forwards in time and not backwards. It would not have been difficult for him then to identify light with one form of these electromagnetic disturbances, whose propagation properties he had deduced entirely from relativity and the most elementary rules concerning the electrostatic field.

It may at this point be worth discussing briefly what the mathematician means by waves. He does not necessarily mean a periodic sinusoidal phenomenon; all he means is something that propagates. As such, one of its essential features is its ability to convey information. It is by means of a wave that the receiver obtains information about what goes on at the transmitter. Changes in the state of the transmitter are propagated as waves; the idea of news is inextricably interwoven with the idea of waves, and this is essential for their understanding.

A second characteristic feature of waves is the transfer of energy. There are many forms of energy transfer, and it may be useful if we think first about the way energy is transferred from a power station to the consumer along the supply wires. The important point here is that the energy put out by the power station is entirely determined by what the consumer actually uses, apart from losses in transmission that are small and could in principle be made negligible. In wave propagation, however, there is a wholly different situation. A radio transmitter radiates power into space, and the amount of power put out is independent of the number and location of the receivers. The amount of energy radiated is quite independent of whether people switch their receivers on or not. This is an amazing property of radiation which in many ways is difficult to understand, for what happens to the energy radiated? It seems to be swallowed by space at large distances. There are good reasons for supposing that the system works only because all radiation is eventually absorbed, though at very great distances, and moreover that this is closely connected with the actual construction of our universe, which is expanding. In the present context the two crucial characteristics of waves may therefore be stated to be that energy is radiated irrespective of the existence and location of receivers, and that waves can convey new information.

GRAVITATION

Gravitation is distinguished from all other forces by the fact, discovered by Galileo, that all bodies fall equally fast. The force of gravitation is proportional to the mass of the body, so that the acceleration of a body is independent of its mass. The full significance of this remarkable phenomenon will be discussed a little later, but when Newton formulated his theory of gravitation he simply took it as axiomatic that the pull of gravitation on any body was proportional to its mass. Furthermore, he assumed the inverse square law of force and that gravitation was a conservative force. Finally, he assumed that the gravitation-producing property of matter, the source strength of the gravitational field, also derived from its mass, a result that in any case follows from his third law concerning action and reaction. A tremendous amount of work has been done on the basis of this Newtonian formulation of gravitational theory. The solar system has given vast scope for applying Newton's theory of gravitation and for checking it over long periods of time in the minutest detail.

Sixty years ago it could be said that Newton's theory of gravitation was the best established and soundest in physical science. Then, quite suddenly, a new theory of gravitation was advanced, Einstein's general theory of relativity, and belief in the absolute validity of Newton's theory was abandoned with little resistance. How could it come about that such a well-established theory was dropped and replaced by another one whose chief merit is that it gives the same results as Newton's theory? The very few and minor points of discrepancy are observationally not too firmly established, and yet the conceptual basis of Newton's theory is regarded as stone dead and general relativity is the theory on which present work is based.

There are two reasons for this eclipse of the Newtonian picture, of which the first is simply special relativity. It was shown in the last section how the basic notions of optics and electromagnetic theory demand that we abandon the Newtonian idea of a uniform even-flowing time and a three-dimensional space and substitute for them a four-dimensional space-time. Newtonian gravitation does not fit into this four-dimensional space-time; it does not have the correct algebraic structure. That this had not become apparent earlier is simply due to the circumstance that all our knowledge of gravitation arises from rather slowly moving bodies—and for such bodies there is almost no difference between Newtonian space and time on the one hand and the space-time of special relativity on the other. Nevertheless, the complete acceptance of special relativity implies that Newtonian theory cannot be fitted into the framework of modern physics and must therefore be abandoned. The choice is now severely limited, as the number of simple relativistic theories is small, and the first suspicion of the existence of gravitational waves arises at this stage. We have seen how the simple laws of electrostatics, including the inverse square law, lead virtually inescapably, within the framework of the space-time of special relativity, to the notion of electromagnetic waves. As soon as one decides that gravitational theory, too, must be relativistic, one suspects that the analogous development should take place here; that is, that while static gravitational fields are well described by the inverse square law, the gravitational fields of fast-moving bodies should give rise to gravitational waves.

The second reason for rejecting Newtonian theory—the principle of Galileo—serves to determine which of the several possible relativistic theories of gravitation should be accepted. In Newtonian theory, the fact that gravitational force is proportional to the mass of the body on which it is acting is considered no more significant than, say, that electrostatic forces should be proportional to the charge of the body on which they are acting. However, this view of the association of mass with gravitational force is rather shallow. Bodies can have any charge we please to give them, but the mass of a body is inalienable. There is no such thing as a body without mass, though there are plenty of bodies without charge. This very directly affects Newton's first law, the very basis of Newtonian dynamics, which states that a body on which no forces are acting moves in a straight line with constant velocity. For what are the chances of testing this law? Suppose we wish to experiment by observing the motion of a suitable body. In order to test the first law we have to remove all the forces acting on it. First, therefore, the physicist carries out an inspection to make sure there are no ropes pulling the body, then he checks that there is no friction between it and its surroundings. Next, he makes certain that it is without charge, so that it is not pulled by electric fields, and so on. In this way he can remove all the forces acting on the body, or rather all the forces with one exception—gravitation. He cannot relieve the body of its mass, because this can no more be separated from it than a man from his shadow.

Newton thought he could get round this difficulty by specifying that the body to be tested should be sufficiently far from any other bodies, but this would scarcely serve the purpose. Not only should the standard motion, which is what the first law prescribes, be available at every point, but it is now known that the universe is everywhere populated with matter, so that no sufficiently isolated region exists. We have, therefore, to conclude that the Newtonian standard is not attainable. The basic premise of the Newtonian standard is that if a body follows the standard motion, then there is no purpose in looking for a force causing it. It is only when the motion of the body deviates from the standard that one has to look for a force. However, this sort of argument fails if the Newtonian standard is not an attainable, testable standard. The only way out is to take as the standard motion of matter its motion under gravitation; in other words, to abandon the whole notion of gravitation as a force. A force is something that causes matter to deviate from its standard motion, the standard being its motion under gravitation and inertia. Gravitation and inertia now become, in some sense, one; they provide the rails on which matter runs when there are no forces present. Naturally, in order to include all the complicated phenomena of gravitation, the new standard motion must be infinitely more complex than motion in a straight line under constant velocity, the Newtonian standard. Nevertheless, a far more satisfactory situation has now been reached, in which due account is taken of the fact that gravitation makes all bodies move equally, and in which the new standard is available.

What remains of the gravitational field when we regard it as the standard? The only property that we can work with now is its inhomogeneity, the fact that neighbouring particles do not fall in precisely the same way, because the gravitational field varies from point to point. A study of the relative acceleration of neighbouring particles now serves to describe the gravitational field, and this can readily be done in a relativistic fashion. Moreover, the form of the relativistic theory of gravitation has now been fixed; it must be such that gravitation reproduces the effects that inertia can produce, because inertia and gravitation are now regarded as one. It will be recalled that in Newtonian theory, when a non-inertial system of reference was used, fictitious forces such as centrifugal force arose. It follows that the description of the gravitational field must correspond to the complexity of motions that can be produced by using arbitrary systems of reference and considering the fictitious inertial forces then arising. On this basis it turns out that gravitation is described by a symmetric tensor in space-time, that is, an entity with ten independent components. This tensor has in fact the special significance that it describes the nature of space-time itself. Space is naturally measured out by using freely moving particles and light rays, the latter being a limiting form of freely moving particles. Therefore, the new standard motion serves to describe space-time itself. It is measured in terms not of straight lines traced with constant velocity, but in terms of the free motion of matter.

This is the foundation of Einstein's general theory of relativity, and it can be seen how immensely compelling it is. Its basis is that all bodies fall equally fast, that being affected by gravitation is an inalienable property of matter. It is, from the start, conceived in space-time, and therefore it fits the notions of electromagnetism and special relativity; consequently it is the natural successor to Newton's theory. There the equation that determines the way in which matter produces the gravitational field consists of a certain linear combination of relative accelerations of neighbouring particles, the 'Laplacian', being equated to the density of matter. A similar procedure in general relativity yields the so-called field equations, which again use linear combinations of the coefficients determining the relative accelerations of neighbouring particles, and equate these to the material content of space. Although this procedure is perhaps not unique, the argument is so compelling that one has little doubt that general relativity is by far the best relativistic formulation of gravitation that also takes account of the fact that all bodies fall equally fast. Moreover, it is shown without great difficulty that, for slowly moving bodies, general relativity gives almost the same answer as Newtonian theory; the discrepancies are very small. One of them concerns an extremely small effect on the motion of the planet Mercury, where Newtonian theory did not quite fit the facts, and there general relativity does. Two other discrepancies are concerned with the propagation of light. One of them can be inferred from very basic principles but is difficult to observe experimentally, though this has been achieved quite recently. The other concerns the effect of gravitation on light rays, and here again the observational results are in fair agreement with relativity, which demands only a very small effect.

To recapitulate, we are forced to abandon

Newtonian gravitation because it is not relativistic. If we try to formulate a relativistic theory of gravitation we are strongly impressed by the need to take account of the fact that all bodies fall equally fast, and then one is driven, almost inescapably, to Einstein's theory. Experimentally it goes beyond Newton's theory, though not very much so. But the burden of the argument here has been that it is the compulsion coming from other branches of physics, more than the agreement with those details of observation where there is a difference between Newton and Einstein, that forces one to adopt general relativity. We are driven to it by the principle of the unity of physics that demands that electromagnetism and optics should not exist in a different sort of space-time from that of gravitation.

THE QUANTUM THEORY

Only a little need be said here about the important understanding that has been gained in the last fifty years of the structure of atoms and nuclei, and of the behaviour of sub-atomic particles. It is well known that the physics governing these differs from that for larger bodies, because, fundamentally, the effect of the means of observation on what is observed can no longer be neglected. When dealing with the smallest entities, one cannot use tools finer than the entities themselves. This introduces indeterminacies and uncertainties that are formalized in quantum theory. The first step here was the quantization of dynamics (quantum mechanics), and the first application was to the motion of electrons in the field of a charged nucleus. This field is electromagnetic but it is effectively only electrostatic. By a fundamental property of quantum theory the electrostatic field itself need not be quantized, essentially because it cannot convey news; it is treated simply as an aura surrounding the central nucleus. At this early stage of the development of the theory it was therefore only in the response of bodies to known forces that allowance had to be taken of the indeterminacies arising from the restrictions on observability. Quantum mechanics in the strictest sense is sufficient for the discussion of the basic features of atomic structures, because there the field is of the special electrostatic type, obeying the inverse square law, that does not need to be quantized. This step, quantization, was only later taken for the electromagnetic radiation field, where it led to grave complications and difficulties that have not been overcome until quite recently.

GRAVITATIONAL WAVES

It is a fundamental demand of quantum theory that there should, beyond a certain degree of accuracy, be no way of ascertaining simultaneously both the position and the momentum of a particle. The gravitational field due to a body is of appreciable strength only if its mass is quite large. On the other hand, the gravitational effect is additive, that is, one half of a mass produces one half of the gravitational effect of the whole mass. If this rule is taken to its limit—and there exists no evidence that it may not be—then there should even be a gravitational field due to a single sub-atomic particle, such as a proton, or neutron, or even electron. Of course, these gravitational fields are incredibly weak, because the masses of the particles are so small, and they are completely hidden by the electrical fields which are immensely stronger, the factor being 10^{39}. Nevertheless, if it is in principle impossible to measure simultaneously both the position and the momentum of the particle, then it must be impossible to find these quantities even by an unbelievably exact determination of the gravitational field. The gravitational field, too, must therefore be quantized to accord with the rest of physics. Once again it is the unity of physics that makes a demand, although so far as actual measurement is concerned we are very far indeed from being able to test the need for it. Logically, however, this does not affect the necessity for quantizing the gravitational field, which must be done to ensure that the same limitations of observability of minute particles apply to gravitational as to other experiments. This difficult process need not be carried out, however, if the gravitational field is simply an aura surrounding the particle. Just as in quantum theory it is not necessary to quantize the inverse-square-law electrostatic field of the central particle, so it would probably not be necessary to quantize the gravitational field if this were purely an inverse-square-law field. If, however, there exists in gravitation a radiation field, then this must be quantized just as the electromagnetic radiation field had to be. Thus we are driven from two different directions to ask whether gravitational waves exist: from the problem of the analogy with electromagnetism, where the application of relativity led from static considerations to wave fields, and from quantum theory, which demands the quantization of the gravitational field if there are such things as gravitational waves.

Finally, there is just ordinary curiosity. We have had a satisfactory theory of the gravitational field, Einstein's theory, for over forty years. It is time we knew what this theory says about the gravitational field of rapidly accelerating particles, even if such a field cannot be observed in nature for what appears to be a quite accidental reason, namely, the absence of sufficiently rapidly accelerated objects of large mass. Thus the demand for unity in physics makes the problem of the existence of these waves interesting, although their experimental verification is unlikely for quite some time.

Having posed the question, one is surprised that it is difficult to get the answer, but the reasons for this are largely mathematical. The equations describing general relativity are, in all but the simplest applications, exceedingly complex and difficult to unravel. Moreover, there is a physical pecularity of their structure. Owing to the unification of inertia and gravitation they allow us to use any system of reference, and it is not easy to specify a system of reference sufficiently without violating the flexibility of space-time that is such an essential feature of the theory. Various methods of approximation have been used, and some of these, at least, are rather suspect. However, considerable progress has been made in recent years and a number of results have been derived.

Firstly, how can gravitational waves be generated? The simplest idea that comes into one's mind concerns the propagation of gravitational changes and could be put into the form of a question. If the Sun suddenly ceased to exist, how soon afterwards would the Earth leave its orbit? However, this is a nonsensical question. The laws of gravitation involve the law of conservation of matter, and so, within the context of gravitational theory, the Sun cannot cease to exist. Next, one might consider what would happen if, instead of ceasing to exist, the Sun suddenly moved off at right angles to the plane of the Earth's orbit. How long would it be before these effects began to show themselves? Again the question is nonsensical, because the laws of gravitation imply the conservation of momentum, and therefore the Sun cannot go off sideways. There are, however, changes that are not forbidden. The Sun could suddenly develop into a prolate or oblate spheroid, and such changes would undoubtedly affect the Earth's orbit. The Sun could explode into two parts going off in opposite directions. These are perfectly reasonable possibilities, and any theory should, at least in principle, clearly state what would then happen.

It turns out that the velocity of propagation of gravitational information is the velocity of light; since this is built into the structure of space-time, the result is in no way surprising. Throughout physics the transport of information is inextricably linked with the transport of energy, though no simple rule exists to determine how much energy is required to convey a given amount of information. To study the question posed above, the relativist considers a body surrounded by empty space and asks how changes in the shape of the body affect its mass. Changes of shape imply changes in the gravitational field, and thus necessarily also imply propagation of information, while the mass of a body is the most convenient measure of its energy. Crudely put, the question is whether changes of shape radiate away any of the mass of the body.

To formulate this question precisely, it is not enough to require the space surrounding the body to be empty, a phrase which in relativity excludes not only matter, but all other fields except gravitational ones. However, if gravitational waves carrying energy exist, such waves might not only be produced by the body and travel outwards from it, but might also be coming inwards from great distances towards the body and so bring energy to it. To exclude such incoming waves, an 'outgoing radiation condition' has to be applied in addition to the demand that space be empty. This condition is closely associated with the application of the principle of causality, namely that the effects of changes of shape should occur only after the changes. For the outgoing radiation condition also prohibits incoming waves specifically arranged to conform to the changes of shape of the body, that might be taken to be waves due to these changes of shape but propagated backwards in time. Thus the condition not only prevents an indiscriminate inflow of energy but a violation of our notions of causality. The formulation of a suitable condition of this type is not easy, and it is gratifying that three different formulations have recently been shown to be equivalent, at least in a case of special importance.

An associated question concerns the origin of the news propagated in the outgoing waves. How can an isolated body come to change shape, especially when it has been isolated for so long that the surrounding gravitational field contains no memory of the body having been put there? Note that the change should in some way be unexpected, so that the waves really convey news. The simplest device would seem to be a time-bomb with a detonator that takes a very long time to go off and is sufficiently defective for its behaviour to be unpredictable. In this way all the conditions of the problem can be satisfied.

The mathematical solution is as yet deplorably incomplete. It has not proved possible to do more than develop some methods of approximation whose validity is not entirely beyond doubt. One of these methods is extremely lengthy and cumbersome, and is so far confined in application to a special case. In another, also of great length, there is some doubt about the unambiguous formulation of the conditions to be satisfied at great distances. A third, less lengthy than the others, confines itself entirely to consideration of the field far from the source. In this last method there is no serious difficulty about satisfying all the conditions at great distances from the source, but one cannot discover which specific distant field is produced by which type of motion of the source.

Nevertheless, some answers have emerged unambiguously, and at least some further questions can be formulated. Thus it is known that if a system that is initially static undergoes some changes of shape and eventually becomes static again, then its mass in the final state is less than its mass in the initial state by an amount proportional to the square of the rate of change of the field, summed through the period of change. The restriction to initially and finally static states is made because only static fields are fully understood and admit of an unambiguous definition of the mass of the system. Thus it has been established that a system can radiate away some of its mass, though this is only a second-order concomitant of changes of shape and not a primary cause of radiation as in the hypothetical case of the Sun's disappearing.

The structure of a gravitational wave must be wholly transverse. Thus, if a wave of limited duration hits a swarm of particles previously at rest relative to each other, then after the passage of the wave there is no relative motion of the particles parallel to the direction of propagation. In one transverse direction the particles will be separating, and in another, at right angles to the first, they will be approaching each other. At large distances from the source a limited section of the wave is almost plane, and plane gravitational waves are well understood. In spite of the complexity of the equations, they travel without change of shape or dispersion. In this, they are like electromagnetic waves (which, however, can be superposed linearly, whereas this cannot be done with gravitational waves) and unlike sound waves, which sharpen up to form shock waves. This constancy of shape is no doubt associated with the transverse nature of the waves.

The construction of a gravitational-wave receiver is also understood, at least in principle. Suppose two massive spheres can slide on a rough stick but are attached at separate positions by springs. Then if a wave travels at right angles to the stick it tends to make the spheres move in relation to one another, as with the particles of the swarm referred to before. This is bound to lead to relative motion between the spheres and the stick, and thus to frictional heating. Some of the energy of the wave has been converted into heat, which is localized energy, and thus the wave has been received. The very response of the receiving system leads to re-radiation, and therefore only a limited amount of energy can be captured by one receiver from a given wave. To make this amount as large as possible, the motion of the spheres must be neither too small nor too large, and thus the concepts of impedance and matching apply in the gravitational as in the electromagnetic case.

A fascinating question arises from the discovery that the distant gravitational field can be changing in a specific, fairly wide, class of ways without there being any radiation. No such change can begin or end without the presence of radiation. What motions of the source can be responsible for such non-radiative changes of the field? Present methods give little indication, but each of two contradictory arguments has some force. According to one, it is lack of acceleration of the particles constituting the source that leads to absence of radiation in spite of source motions. This fits in both with electromagnetic radiation and with the first term of a particular method of approximation, but neither of these arguments is necessarily relevant. The opposing argument stresses that lack of acceleration is irrelevant in gravitational theory because the natural orbits of particles are not the unaccelerated ones, but those of free fall under gravitation. On this view, it is a dust cloud that is likely to be the source of non-radiative changes, because in such a cloud each particle moves freely. But if a dust cloud does not radiate then it does not seem plausible that a passive solid body could; a time-bomb, the prototype of a radiator, is highly active by contrast. Put more mathematically, the relation between pressure and density is time-independent in a dust cloud or in an elastic solid body, but time-dependent in a time-bomb. If there is any form of internal friction—imperfect elasticity, viscosity, and so on—this relation depends on the sense of time, and the processes occurring are irreversible. It is therefore tempting to suggest that the important class of time-dependent but non-radiative fields at large distances are produced entirely by reversible changes in the source of the field, but so far this conjecture is wholly unproved. If it is correct, then there is far less gravitational radiation in the universe than there would otherwise be. The solar system would radiate about one kilowatt in gravitational waves if these were produced by all accelerations, but merely a fraction of a milliwatt if only irreversible processes, such as tidal friction, led to radiation. The question is fascinating, but there does not seem to be any easy way of answering it.

Finally, brief reference may be made to another unsolved problem, that of the applicability of Huygens's principle to gravitational waves. All equations of the wave type prohibit the transmission of any influence at speeds exceeding the characteristic velocity, but only some imply that all influences necessarily travel at just this speed. Thus the ordinary equation of small-amplitude sound-waves shows that in plane and spherical sound-waves the entire influence travels at the speed of sound; for example, after the passage of a brief pulse the medium is wholly at rest again. In the case of cylindrical sound-waves, however, while particles rest in front of the pulse, behind it there is disturbance, because some of the energy travels at speeds lower than the velocity of sound. Thus Huygens's principle, according to which the entire disturbance travels with the characteristic velocity, is satisfied for plane and for spherical sound waves but not for cylindrical ones. In the case of gravitational waves it is known that Huygens's principle applies to plane waves and does not apply to cylindrical ones, but the physically most important case of spherical waves (radiation from an isolated system) is quite obscure, for purely mathematical reasons. A solution of this problem would clear up many other uncertainties in the subject.

We are still quite some way from really understanding gravitational waves and even further from quantizing them; but some progress has been made, and thus we are nearer to being able to demonstrate the unity of physics.

BIBLIOGRAPHY

[1] EINSTEIN, A. 'Meaning of Relativity' (fourth edition). Princeton University Press, Princeton. 1950.

[2] *Idem.* 'Principle of Relativity' (fifteenth edition). Methuen & Co. Ltd, London. 1954.

[3] RUSSELL, Bertrand. 'A.B.C. of Relativity'. George Allen and Unwin Ltd, London. 1958.

[4] COULSON, C. A. 'Waves' (seventh edition). Oliver and Boyd Ltd, Edinburgh. 1955.

[5] LANDAU, L. D. and RUMER, G. B. 'What is Relativity?' (translated by N. Kemmer). Oliver and Boyd Ltd, Edinburgh. 1960.

The green flash and kindred phenomena

D. J. K. O'CONNELL

When the Sun is near the horizon it shows phenomena not seen at any other time. Perhaps the most dramatic of these is the transitory green flash that may sometimes be seen. Once suspected of being a physiological phenomenon, colour photography has shown the green flash to be objective: its appearance can be explained by the ordinary laws of optics. The phenomena depend not only on conditions near the ground but on those in the upper atmosphere, and may be helpful in investigating the latter.

Most voyagers in tropical seas have memories of glorious sunsets, with ever-changing colours spread over a vast expanse of sky. Some may have been fortunate enough to glimpse a fleeting effect when, at the moment of sunset, the last tip of the sun turns a brilliant green. This is the green flash, or the green ray as it is sometimes called. The first seems the more appropriate name, for what is seen is a thin horizontal green strip. A succession of green flashes may, however, appear to the naked eye as a ray shooting up from the sun (figure 2).

One need not go to the tropics to see the green flash, for it can be seen in any latitude, and over the land as well as over the sea. It is also sometimes visible at sunrise (figure 5). The flash is not necessarily green, but may even be blue or violet; Lord Kelvin [1] described a blue flash that he had seen in 1899 as the sun rose over Mont Blanc. Some years previously he had seen the green flash at sunset, and he wanted 'to see the earliest instantaneous light through very clear air, and find whether it was perceptibly blue'. He continues: 'I therefore resolved to watch for an hour till sunrise, and was amply rewarded by all the splendours I saw. . . . In an instant I saw a blue light against the sky on the southern profile of Mont Blanc, which in less than one-twentieth of a second became dazzlingly white. . . .'

The earliest known published account of the phenomenon is that given by D. Winstanley [2] in 1873. 'The green ray . . . begins at the points or cusps of the visible segment of the sun, and when the setting is nearly complete, extends from both cusps to the central space between, where it produces the momentary and intense spark of coloured light visible to the unaided eye'. This is an excellent description of the way in which the green flash has often been seen at my observatory. Winstanley quotes from a letter of Joule, describing his own observations of the flash and still earlier ones by Baxendell. The earliest dated observation that I have been able to trace was made by Swan in 1865 [3]: from the Righi at sunrise he saw a dazzling emerald-green flash over a distant mountain. He did not, however, publish this observation until 1883, perhaps after reading Jules Verne's romance *Le rayon vert*, which appeared in 1882 and which seems to have aroused widespread interest in the subject. At any rate, a great deal has been written since [4–6].

Opinion has been divided about the nature of the green flash. Many held it to be merely a subjective phenomenon, resulting from fatigue of the retina causing the complementary colour to be seen after the eye had been dazzled by the reddish rays of the setting sun. Rather curiously, this was the opinion of Swan after making the observation mentioned above, although it would seem evident that a green flash at sunrise cannot be the result of retinal fatigue; Swan blamed the reddish colour of the dawn sky.

On the other hand, many scientists believe that the green flash is an objective phenomenon [7]. I have myself been firmly of that opinion ever since I first saw the flash, at sunset over the Mediterranean, from the window of my office at Castel Gandolfo, and it seemed worth while trying to settle the controversy by photographing the flash in colour. The Vatican Observatory at Castel Gandolfo is well situated for such an experiment: there is an unobstructed view across the Roman Campagna to the Mediterranean, and the telescopes can be pointed right down to the horizon fifty miles away. It proved no easy task, but—thanks to the skill and patience of C. Treusch, S.J., the instrument-maker of the Observatory—many successful photographs have been taken not only of the green flash at sunset over sea and over land, and at sunrise, but also of various kindred phenomena [6].

THEORY OF THE GREEN FLASH

The main factors that are responsible for the green flash are the dispersion and absorption of the sun's light in the earth's atmosphere. The velocity of light decreases as the density of the air increases, so that a ray of light is refracted in its passage through the atmosphere, unless its source is exactly in the zenith. The ray is tilted towards the vertical, as compared with its original direction and consequently the sun or a star appears higher in the sky than it really is. The refraction increases with increasing distance from the zenith, and rays of shorter wavelengths (violet, blue, green) are bent more than those of longer wavelength. Thus the net result is that the image of a star is drawn out into a spectrum, with blue or green above and red below, the length of the spectrum increasing with decreasing altitude. This effect is shown very clearly in a photograph of Venus taken when it was near the horizon (figure 7): a series of images of the planet in different colours can be seen. The sun, when near the horizon, has a narrow green rim above and a red rim below, neither visible to the naked eye (figure 4).

The air not only refracts light but absorbs it. As a star sinks, its light passes through an ever-increasing thickness of air; the absorption increases and the star becomes steadily fainter. The absorption is not, however, uniform over the whole spectrum. Certain normal constituents of the atmosphere, such as water vapour, oxygen, and ozone, absorb some wavelengths more than others. The result of this is that certain colours fade more rapidly than others as a star nears the horizon. The absorption bands due to water vapour, often called Brewster's bands, after their discoverer, affect the orange and yellow rays very strongly, and sometimes, when the sun is setting, the orange and yellow light almost or quite disappears, so that there is a gap in the spectrum and an abrupt transition from red, or orange-red, to green.

Another factor that influences the flash is the scattering of sunlight by the molecules of the air. Because of this scattering, which also explains why the sky is blue, the violet and blue content of the sun's light grows fainter and fainter as the sun nears the horizon. When the rest of the sun has passed below the horizon, only a narrow green strip is left, the blue and violet being no longer visible. Under these conditions a green flash may be seen at the last moment of sunset. The flash may even appear completely detached from the sun's disc (figure 3) and float above the horizon after the sun has disappeared. At high altitudes, and when the air is very clear, the blue may still be visible at sunset and the flash may be blue-green, or pure blue, or, much more rarely, violet.

Dust or haze can produce a general absorption of sunlight, as well as selective absorption or scattering of particular wavelengths. In general, dust or haze makes the appearance of a green flash less likely, especially when the sun is much reddened. On the other hand, certain conditions may increase the intensity or duration of the flash, and thus make it easier to see. When the temperature of the air increases with height, what is called an inversion layer is formed in the atmosphere, and it is under such conditions that mirages occur. J. Evershed [8], on a ship in the tropics, frequently saw a brilliant green flash at sunset, and noted that on each occasion a mirage effect was conspicuous, caused by a thin inversion layer near the sea. 'At sunset the last segment of the disappearing limb was . . . reflected and reversed, causing a lenticular shape with the cusps raised about 1′ above the horizon. The green flash occurred when the green-edged cusps coalesced into a single bright patch. . . . It seems to me evident . . . that the mirage layer greatly intensifies the ordinary dispersion effect, by adding the light from the reflected image to the direct image at the moment of setting.' Reflections in inversion layers can be seen in figure 1. Experimental confirmation of the increase in intensity of the green flash due to mirage conditions has been provided by T. S. Jacobsen [9], who photographed the spectrum of the green flash and compared the intensity of the green light in it with that of the green in the spectrum of the low sun.

The duration of the flash depends on the rate at which the sun is rising or setting; this rate varies with the season. The vertical width of the green rim is about ten seconds of arc, which, in middle latitudes, would take only a fraction of a second to pass below the horizon. Atmospheric conditions may increase the width of the green rim, with a corresponding increase in the duration of the flash. At Castel Gandolfo the flash has often been seen to last one or two seconds or more, while our measures of the width of the green rim have reached, and at times exceeded, 30 seconds of arc.

The sun sinks more slowly, and the green flash lasts longer, the nearer one is to the poles. At Hammerfest (79°N), at one period in the year, the flash may last as long as 14 minutes, seven minutes during the slow sinking of the sun below the horizon and another seven minutes at sunrise,

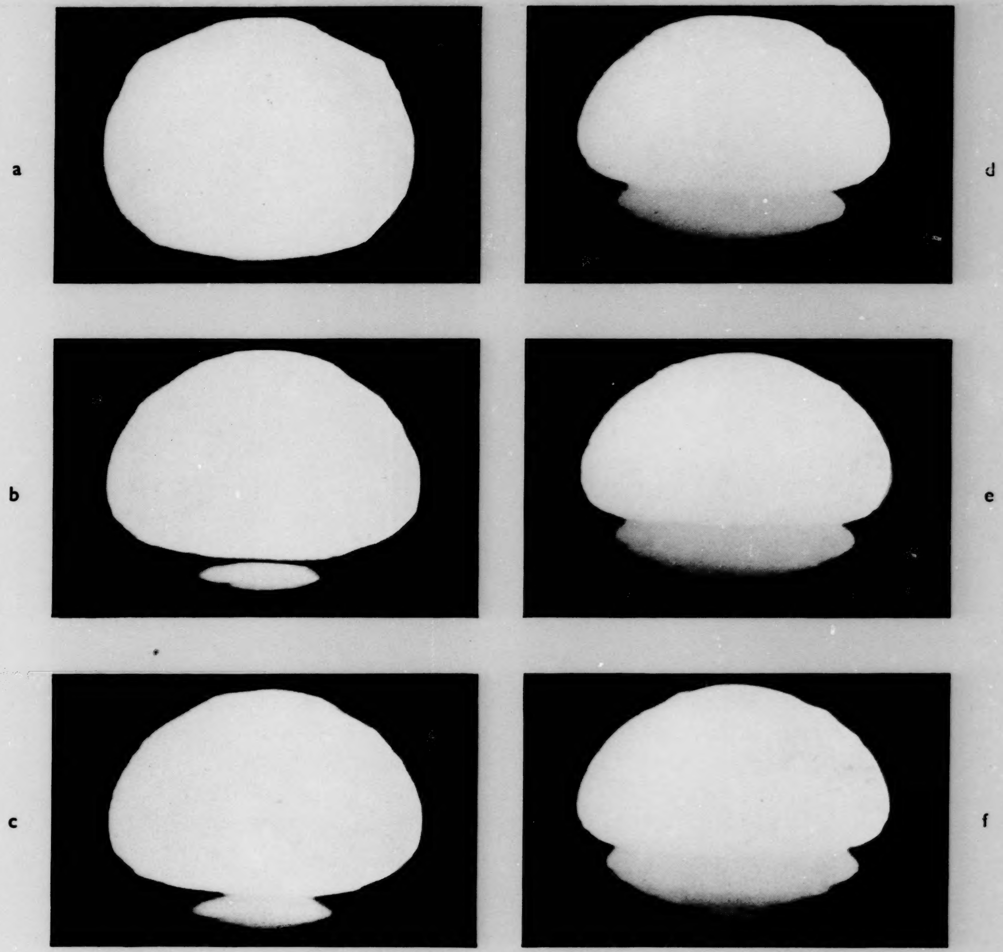


FIGURE 1 – *This picture was taken with Robot camera (aperture 7 cm, focal length 100 cm) on 15th September 1956, to show what can be achieved with a comparatively short focus. The distortions of the sun's disc are well shown, but it is not possible to obtain, with a focal length of only one metre, reproducible colour photos of the green flash. The effects of a 'blind strip' are, however, well shown. A small red flash could be seen with the naked eye in* e. *Some three minutes later a green flash appeared at the same point.*

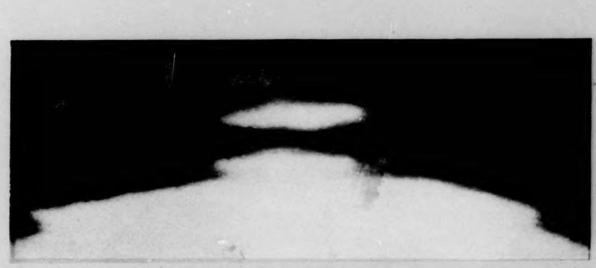


FIGURE 2 – *Multiple green flash at sunrise. A series of green flashes shooting out of the sun's rim, as seen here, can, when seen with the naked eye, give the appearance of a green 'ray'. Taken with 40 cm refractor.*

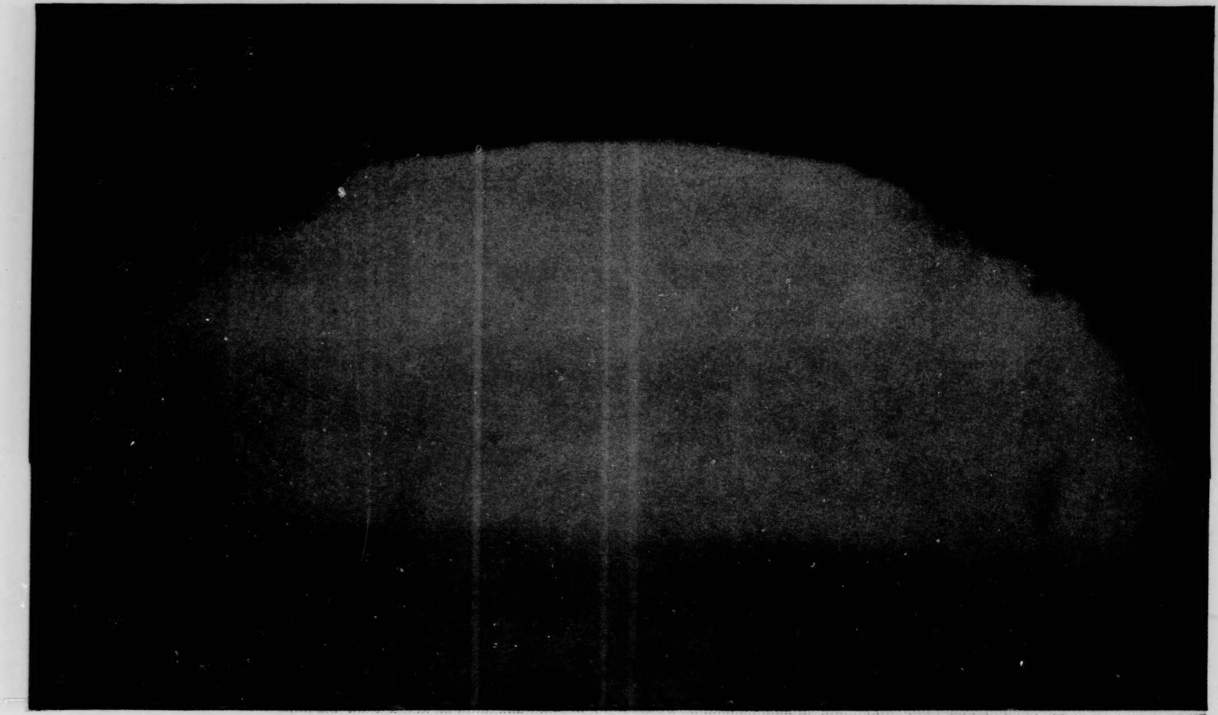


FIGURE 3 – *Green flash, shortly before sunset over the Mediterranean. (Taken on Kodachrome with 60 cm reflector.)*

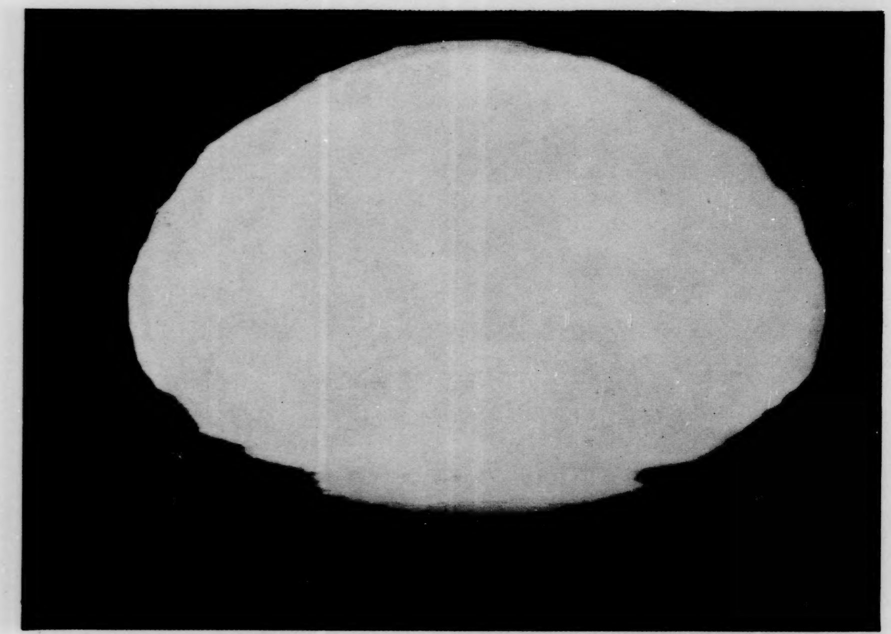


FIGURE 4 – *Green and red rim, with red flash, which lasted about one second, 13th August 1954. Below the red flash can be seen the horizon, a dark red line caused by the reflection of the red rim in the sea. (Taken on Ferraniacolour Reversal with 60 cm reflector, stopped down to 10 cm.)*

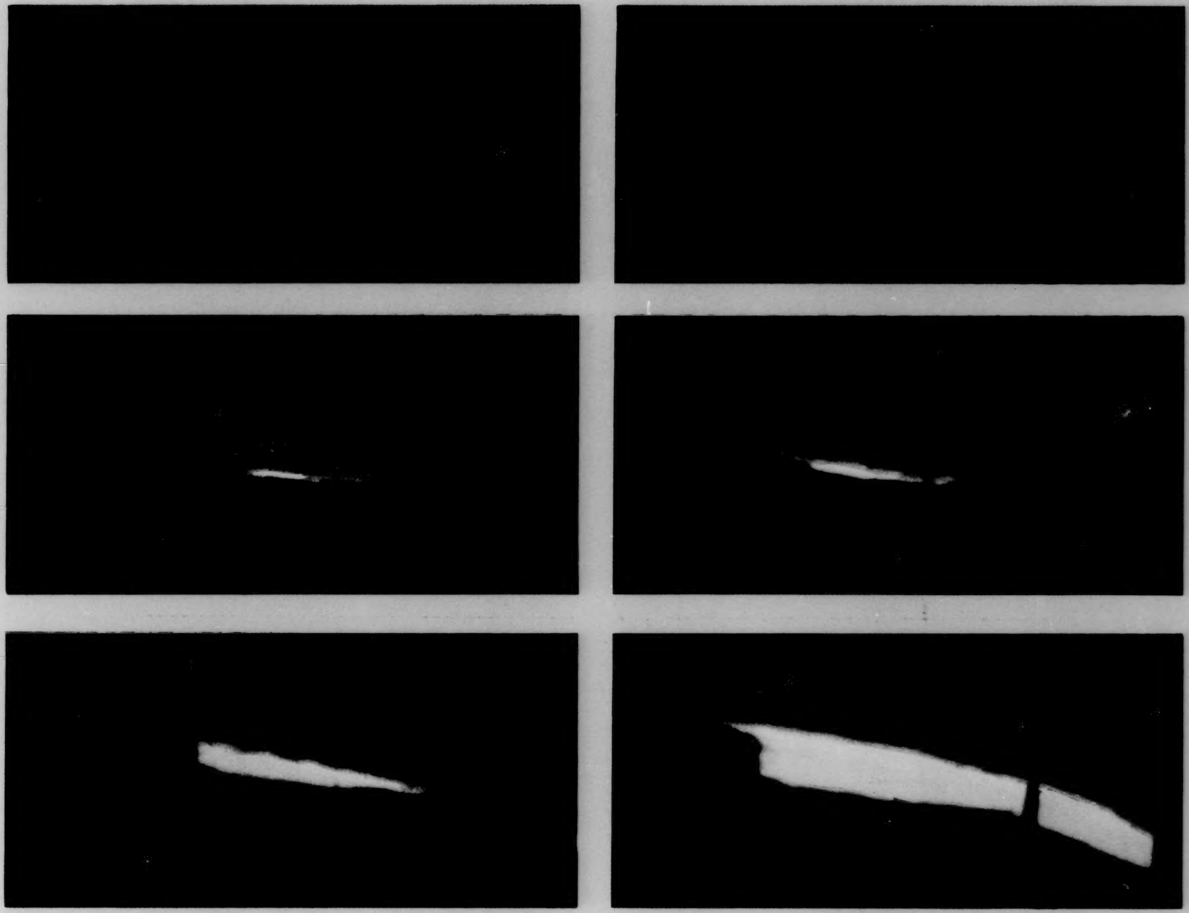


FIGURE 5 – *Green flash seen at sunrise behind medieval castle at Rocca Priora, 16th July 1956. (Taken on Ektachrome with Robot camera, 3 pictures per second, 1/250th sec. exposure, with 40 cm lens stopped down to 10 cm.)*

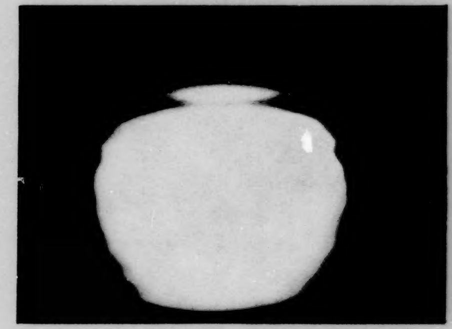


FIGURE 6 – *Artificial blue and red flashes. (Taken with Zeiss-Tessar on Ektachrome B, as described in text.)*

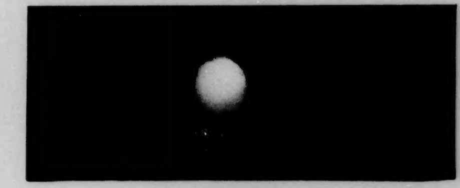


FIGURE 7 – *Venus setting. (Taken on Ektachrome Daylight with 40 cm refractor, exposure 10 sec. Magnification ×23·7.)*

following immediately on sunset [10]. A further increase is obtained when the sun at rising or setting follows the slope of a mountain. M. Minnaert [11] was able to keep the green flash in view by moving along a slope at the appropriate speed. The longest duration on record is that reported by Byrd's Antarctic expedition at Little America (78°S) in 1929, when the sun grazed the irregular horizon of the barrier ice and the green flash was seen on and off for 35 minutes. There was a strong temperature inversion near the ice, which no doubt increased the intensity.

As explained above, refraction causes the sun to appear higher in the sky than it really is. At sunset, the sun's image is raised so much that it is seen for an appreciable time when it is really below the horizon; in other words, sunset is delayed by refraction, just as sunrise is earlier for the same reason. At sunset over the Mediterranean, as seen from Castel Gandolfo, the sun is raised by refraction on the average about 1·3°, about 2·5 times the sun's diameter, and sunset is delayed by several minutes. The degree of refraction varies, depending on atmospheric conditions; it may happen that the refraction is appreciably greater than normal. In that case, sunset, besides being delayed more, lasts longer than usual, and the duration of the flash is also increased. Remarkable instances of abnormal refraction are on record. The most striking case is that reported by Dutch mariners at Novaya Zemlya in 1597. They saw the sun for fourteen days at a time when it should have been below the horizon, the refraction being more than 4°. This report was studied by Kepler and others, and many came to the conclusion that the observation was impossible and that there must have been some confusion in the dates, owing to the recent Gregorian reform of the calendar. The observations were, however, quite correct, as has been shown recently by S. W. Visser [12], who also points out that Sir Ernest Shackleton noted a similar phenomenon in the Weddell Sea during his last Antarctic expedition in 1915.

Since the formation of the green flash may be influenced by so many factors, it is not surprising that its appearance seems quite capricious. R. G. Aitken, for example, saw it at sunset from the Lick Observatory quite regularly for a long period, but at other times looked for it in vain [13]. It has been observed from many parts of the earth, but one is more likely to see it in tropical seas, or on high mountains, or in the clear air of Egypt. There is some reason to think that it was familiar to the ancient Egyptians [14]. It can indeed be seen very frequently in Egypt at sunset, and even more easily at sunrise, for the air is then particularly clear and free from dust.

The green flash is so extremely narrow that it is below the resolving power of the human eye; it can be seen with the naked eye only when the sun's disc has disappeared and the flash remains for an instant above the horizon. It is much easier to see if one uses binoculars or a small telescope, though one must, of course, be extremely careful not to injure the eyes when so doing. A neutral filter between the eyepiece and the eye can be used to reduce the intensity of the light without altering its colour; the purpose of using an optical aid is to increase the magnification. One can often see the green flash quite well with small binoculars, even when it is not visible to the naked eye. If, however, one wishes to distinguish the many fine details, such as have often been photographed here, it is necessary to have a telescope of fairly long focal length.

THE VATICAN OBSERVATORY PHOTOGRAPHS

Many have tried to photograph the green flash in colour with a short-focus camera. When the attempt failed, they concluded that the phenomenon was subjective. The Vatican Observatory photographs were taken with the Zeiss 60 cm reflector (focal length 240 cm) or with the Zeiss 40 cm visual refractor (focal length 600 cm). At first we used the reflector, in order to make sure that no colour effects were introduced by the optics. Later experiments showed that pictures taken with the refractor were free from spurious colour effects, and this telescope was used for most of the later photographs.

The first colour photographs were taken on Kodachrome film; it was a tedious business, for at that time the film had to be sent overseas for processing. As soon as Ektachrome film became available it was used almost exclusively, for it has the very great advantage that it can be developed on the spot. Much experiment was necessary to find the correct exposure, for the intensity of the light is changing and the effects to be photographed are so very fugitive. Correct exposure-time means much more than just recording the phenomena, for every effort was made to ensure that the colours in the film corresponded closely to those seen visually at the same time in the telescope.

During most of the year the sun, viewed from Castel Gandolfo, sets over the Mediterranean, and most of the sunset pictures of the green flash were taken then. Successful pictures of the flash have also been taken at sunset over land, when, during part of the summer, the sun sets behind the Tolfa mountains, sixty miles away [15]. Recording the green flash at sunrise (figure 5) presents a much more difficult task. The telescope must be pointed precisely where the flash may appear, and one must estimate beforehand the exact instant of appearance and the correct exposure. To make matters more difficult still, our eastern horizon, the Alban Hills, is rather high and close, and most of the horizon line is covered by forest. Nevertheless, pictures of the flash have been obtained. Interesting photographs have also been obtained of the blue and green rim of the sun rising behind the forests on the ridge of the hills.

There are many interesting phenomena to be observed and photographed when the sun is near the horizon. Its shape is distorted, the lower rim being raised by refraction more than the upper, so that the disc appears compressed vertically. Reflection in the sea, or in an inversion layer, gives it the form of a capital omega (Ω). There are often layers of discontinuity in the atmosphere, where the rate of decrease of the density of the air with increasing height changes abruptly. These layers cause notches to appear on the sun's rim (figure 4). An inversion layer, that is a layer of air warmer than the surface of the earth beneath it, sometimes causes a 'blind strip' which cuts the sun's disc in two (figure 1; see also [11]). At the lower edge of the blind strip there sometimes appears a red flash, and soon afterwards when the upper rim of the sun is about the same height above the horizon, a green flash may appear. Such red and green flashes have frequently been seen and photographed here (figures 1 and 4). Sometimes green strips form at the upper edge of the sun and gradually become detached, to form a green flash even before the sun sets (figure 3). At times a succession of such detached wisps or flashes have been seen here and photographs have been taken. Figure 2, taken at sunrise, shows three such green flashes, one above the other. Much more rarely seen than the green flash is a red flash below the sun. The red flash has been photographed here several times (figure 4). Venus presents an interesting sight when setting, as can be seen from figure 7.

It is interesting to study the effects of scintillation on the sun's rim: the changes are so rapid that it requires a cine camera to follow them. Successful photographs of scintillation effects, and of the development of detached green strips, have been made with a cine camera (without lens, of course), attached to the 40 cm refractor. The long focal length of this telescope makes possible the recording of very fine details that would be lost with a shorter instrument.

Figure 6 shows the results of an experiment to produce artificial blue and red flashes. In this, a glass mirror was placed on the ground under a flint-glass objective prism of 60 cm diameter and 8° refracting angle. The light from a ceiling lamp, an opal sphere of 20 cm diameter, passed through the prism, was reflected by the mirror and passed back through the prism to a camera; this simulated very well the image of the sun when near the horizon. To reproduce the effect of discontinuity layers, three panes of window glass were set above the prism, slightly inclined to the horizontal.

In conclusion, it is worth emphasizing that the appearance of the green flash, and of other phenomena referred to here, is dependent on the state of our atmosphere, not only on the dense lower layers of the air but on conditions at great altitudes. Just now, when men are undertaking their first extra-terrestrial voyages, it is essential to know all that we can about the outer atmosphere. We are indeed learning much by means of rockets and artificial satellites, but there is still a great deal that we do not know about these very tenuous upper regions of the air. It may well be that studies of such phenomena as the green flash may contribute useful information.

REFERENCES

[1] KELVIN, Lord. *Nature, Lond.*, **60**, 411, 1899.

[2] WINSTANLEY, D. *Proc. Manchr. Lit. Phil. Soc.*, **13**, 1, 1873.

[3] SWAN, W. *Nature, Lond.*, **29**, 76, 1883.

[4] FISHER, W. J. *Pop. Astr.*, **29**, 251, 382, 1922.

[5] KUIPER, P. F. 'De Groene Straal', Helder. 1926.

[6] O'CONNELL, D. J. K. 'The Green Flash and Other Low Sun Phenomena'. Vatican Observatory. North Holland Publishing Co., Amsterdam; Interscience Publishers Inc., New York. 1958.

[7] See, for example, DANJON, A. and ROUGIER, G. *Ann. Obs. Strasbourg.*, **1**, 105, 1926.

[8] EVERSHED, J. *Nature, Lond.*, **111**, 13, 1923.

[9] JACOBSEN, T. S. *J. R. astr. Soc. Can.*, **46**, 93, 1952.

[10] GUGLIELMO, G. *R.C. Accad. Lincei*, (5), **25**, 417, 1916.

[11] MINNAERT, M. 'Light and Colour in the Open Air' (2nd ed.), p. 59, Dover Publications, New York. 1954.

[12] VISSER, S. W. *Proc. R. Acad. Sci. Amst., B*, **59**, 375, 1956.

[13] AITKEN, R. G. *Leafl. astr. Soc. Pacif.*, **3**, 177, 1939.

[14] See [6], p. 8.

[15] See [6], plate XI.

Optical rotatory dispersion

CARL DJERASSI

The variation of optical rotation with wavelength has been known for nearly one hundred and fifty years, but its application to organic chemical and biochemical problems started only recently. Measurements of anomalous optical rotatory dispersion have given a powerful new method for the solution of various structural, stereochemical, and kinetic problems. In particular, rotatory-dispersion studies were found to be almost uniquely useful for determining the absolute configuration of organic compounds and in the detection of subtle conformational changes. Some of these recent developments are here discussed.

The organic chemist, especially the investigator of natural products, has always been very ready to adopt new physical methods for the solution of his problems. Today, it is difficult to imagine modern organic chemical research being conducted without the use of ultraviolet, infrared, or nuclear magnetic resonance, spectroscopy, and it is profitable to consider for a moment how these tools became acceptable to the organic chemist. In every case the basic principle was discovered by a physicist or physical chemist, who was usually concerned only with the phenomenon itself, rather than with its application. It was used in the organic chemical laboratory only when instrumentation became sufficiently advanced, or simplified, for a relatively large number of measurements to be made with organic compounds. These measurements were then empirically correlated with some structural feature of the organic molecule, and this invariably led to two developments. One, a rapid acceptance of the method by organic chemists in general, with a consequent enormous increase in the number of measurements, and hence refinements in correlation. The other, a revival or extension of interest in the theoretical aspects, and a more fundamental explanation of the semi-empirical discoveries of the practical chemist.

This has been the history of virtually all physical methods that have found a temporary or a permanent place in organic and biochemical research; the only difference between them has been the time-lag between the initial physical discovery and the first widespread use by the organic chemist. In the field of infrared spectroscopy, for instance, the interval amounted to approximately forty years, but less than a dozen years passed between the discovery of nuclear magnetic resonance and its extensive use in organic chemistry.

HISTORY

The development of optical rotatory dispersion followed this general pattern, except that nearly one hundred and forty years elapsed between the original discovery by J. B. Biot [1] of the change of optical rotation produced by quartz with wavelength and the application of this phenomenon to organic chemical problems [2]. In the interval, considerable research was conducted, largely by physical chemists, and reviews by three of the pioneers, T. H. Lowry, P. A. Levene, and W. Kuhn [3], show on the whole a rather understandable preoccupation with theoretical aspects. During this entire period, up to about 1954, considerably less than one hundred optical rotatory-dispersion curves had been recorded in the ultraviolet region of the spectrum, principally because of the instrumental difficulties. Indeed, between the 1930s, when experimental work by the research groups led by these three men had terminated, and 1955, only a few publications appeared on optical rotatory-dispersion measurements in the ultraviolet, and there are practically no records of the use of this method to solve an organic or biochemical problem. This contrasts with two striking indications of present activity and interest in optical rotatory-dispersion, with special application to organic chemistry and biochemistry. The advances in instrumentation are illustrated by the fact that, in the author's laboratory, approximately three thousand ultraviolet rotatory-dispersion curves were measured in less than eight years. A quantitative measure of the general acceptance of optical rotatory-dispersion as a physical tool is afforded by the fact that, during the past eighteen months, approximately two hundred publications have appeared in which ultraviolet rotatory-dispersion measurements have been employed for the solution of organic chemical or biochemical problems.

GENERAL PRINCIPLES

To appreciate the reasons for this renaissance, a brief discussion of optical rotatory dispersion is required. We may consider a beam of plane-polarized light to be made up of a right and a left circularly polarized component. Although no rotation will be observed if these two components pass through a medium with equal velocities, the plane of polarization of the emergent light will be rotated if the refractive indices of the medium for right and left polarized light are unequal. This dispersive feature is the principal reason for monochromatic polarimetry as, for instance, in the conventional measurement of optical rotation at one wavelength, usually the sodium D line. In addition to unequal velocity of transmission, there may also occur unequal absorption of left and right circularly polarized light. The emergent light will then be elliptically polarized; this absorptive effect is known as circular dichroism. The combination of circular dichroism and optical rotation is named the Cotton effect, after its discoverer. It was discovered in 1895. As will be shown, it is the experimental measurement of the Cotton effect, and its correlation with certain structural and stereochemical features, that is chiefly concerned in the application of optical rotatory dispersion in organic chemistry.

In principle, the experimental determination of an optical rotatory-dispersion curve is a simple matter that involves the measurement of the optical rotation of a given substance at different wavelengths, rather than just at the conventional sodium D line. The apparatus consists essentially of a light source; a monochromator producing light of a given wavelength; a polarizer; a cell containing the optically active material in solution; an analyser; and a device for measuring the angle of rotation. The polariser and analyser are usually quartz Rochon prisms, and the angle of rotation is measured visually when working with visible light. However, most organic compounds are colourless, and absorb only in the ultraviolet; consequently, if one wishes to measure the Cotton effect of colourless substances, the determinations must be carried out with ultraviolet light. It is here that experimental difficulties arise. All the earlier work [3] in the ultraviolet region used laborious photographic procedures, and it was only the commercial development of a spectropolarimeter [2] that made possible the relatively rapid and simple measurement of a rotatory-dispersion curve in the ultraviolet. The measurement could be made in an hour or two by this method, whereas the photographic one needed several days.

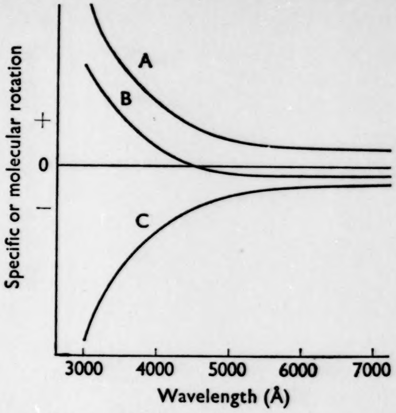


FIGURE 1 – *Typical plain rotatory-dispersion curves* (*A and B, positive; C, negative*).

There are two types of dispersion curves [4]. Figure 1 illustrates some plain curves, that is, dispersion curves that exhibit no 'maxima' or 'minima'; in these the rotation simply increases as one proceeds into the ultraviolet. Curves A and B of figure 1 are called positive plain curves, even though curve B starts out on the negative rotation side in the visible and crosses the zero rotation axis in the ultraviolet; curve C is called a negative plain curve.

An anomalous optical rotatory-dispersion curve is one that exhibits rotational maxima and minima; an idealized example is reproduced in figure 2. Curve A represents a single positive Cotton effect; Curve B is typical of a negative Cotton effect, since the first extremum in going from the visible to the ultraviolet is a rotational minimum. The rotational maxima and minima are called peaks and troughs, so as not to confuse them with ultraviolet absorption maxima and minima. Thus the positive Cotton effect of curve A in figure 2 would be exhibited by an optically active substance with a chromophore absorbing at wavelength λ_0.

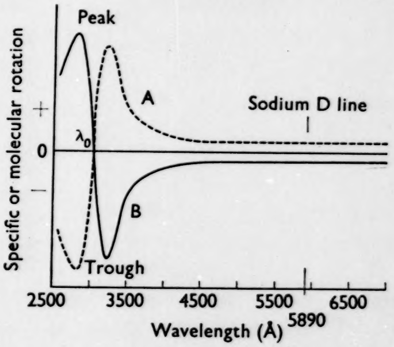


FIGURE 2 – *Typical anomalous rotatory-dispersion curves* (*A, positive Cotton effect; B, negative Cotton effect*).

This represents approximately the midpoint, in terms of wavelength units, between the positions of the peak and the trough of the rotatory-dispersion curve. This intimate relationship between the absorption spectrum and the optical rotatory-dispersion of a given substance will usually indicate whether it will exhibit an anomalous rotatory-dispersion curve in a given wavelength region.

Plain curves of the type shown in figure 1 will be exhibited generally by optically active substances, such as hydrocarbons, alcohols, and carboxylic acids, that do not absorb in an accessible region of the spectrum; an anomalous curve will be shown by a substance containing a chromophore, such as a carbonyl or nitro group, that does absorb in such a region. In reality, every plain curve will eventually become anomalous as it approaches its optically active absorption band; but in many compounds these bands may occur so far down in the ultraviolet, or the compounds may absorb with such intensity, that rotation measurements are not feasible in that spectral region. To illustrate this point, the anomalous rotatory-dispersion curves shown in figure 2 are actually plain curves over the range 7000–4500 Å and become anomalous only as they approach the region of the absorption maximum ($\lambda_0 = 3000$ Å).

For substances whose behaviour is represented by either figure 1 or figure 2, the optical rotations will be much greater in the ultraviolet than in the visible. The historical choice of the sodium D line for monochromatic rotation measurements is therefore a rather poor one. Much smaller amounts of optically active substrate are required for rotation measurements in the ultraviolet than in the visible, and most of the ultraviolet rotatory-dispersion measurements in our laboratory have been conducted with 0·1–1·0 mg of material, a quantity that is usually much too small for rotation measurements at the sodium D line.

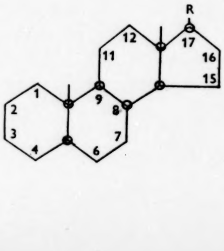
Between 1812 and 1954 only approximately fifty Cotton-effect curves had been measured experimentally [3], and the various chromophores that had been investigated included a few ketones, such as camphor, whose rotatory-dispersion curve was similar to curve A in figure 2. While the availability of this experimental material did not lead to any widespread applications, it nevertheless demonstrated the feasibility of measuring Cotton effects, provided that substances with suitable chromophores were selected. When the first model of a commercially available spectropolarimeter, the Rudolph spectropolarimeter, was delivered to our laboratory, we therefore set out to see if the Cotton effect could be related to structure.

APPLICATIONS TO STEREOCHEMISTRY

Although for stereochemical correlations [5] plain rotatory-dispersion curves are preferable to single rotation measurements at the sodium D line, especially when the curves cross the zero line, they are obviously limited in scope. On the other hand, if variations in the asymmetric environment of a given spectroscopically accessible chromophore should result in Cotton-effect curves of different shape or sign, and if these differences could be related to certain structural or stereo-chemical features of the molecule, a valuable physical tool would be available to the organic chemist.

We selected the carbonyl chromophore for initial experimental examination because it absorbs in an accessible spectral region (2800–3300 Å) with a rather low extinction that would permit rotation measurements through this region of maximal absorption; this is essential if Cotton effects are to be determined experimentally. Furthermore, earlier work [3] on camphor and a few related terpene ketones had shown that such ketones exhibit Cotton-effect curves—in other words, that the intrinsically symmetrical carbonyl chromophore $>C{=}O$ became optically active when placed in an asymmetric environment, although it could not be determined from these few studies whether the Cotton effect would reflect structural changes in its asymmetric surroundings. A further reason for selecting the carbonyl chromophore was the ready availability of many optically active ketones and aldehydes, or at least alcohols from which they could be obtained by oxidation.

With the choice of the chromophore—that is, the absorptive component of the rotatory dispersion—settled, it was necessary to decide upon a reasonable asymmetric framework that would offer sufficient scope for selective alteration to see whether such changes would result in variations in the shape or sign of the Cotton-effect curve. For this purpose we chose a uniquely suitable group of compounds, the steroids, which are characterized by the following skeleton:



The advantages of this group are manifold. All steroids are optically active, of known absolute configuration, and usually of fixed and known conformation. Furthermore, probably several thousand representatives of this class have been prepared during the past thirty years. Finally, most of the steroids possess a chain of seven contiguous asymmetric carbon atoms (circled), so that irrespective of where a carbonyl group is placed in this molecule, it can, at most, be only one carbon atom away from this asymmetric chain. By placing a carbonyl group in the eleven possible locations of the four-ring skeleton, the chromophore will find itself in different asymmetric situations. The incorporation of a trigonal carbon atom into any one of the rings will also cause a displacement of the valency angles, with a concomitant ring-geometry distortion that may make itself felt through varying ranges of this asymmetric backbone. As we had in our laboratory many such steroids, and also a modern spectropolarimeter, we could measure, in the ultraviolet, many hundred rotatory-dispersion curves of closely related ketones. This initial work [6] demonstrated that variations in the shape and sign of the Cotton effects can indeed be related to many important molecular parameters. This prompted us to carry out a much more extensive investigation of the application of rotatory-dispersion measurements in organic chemistry than had been contemplated initially. Much of this work has appeared between 1955 and 1960 in some forty articles from our laboratory, and it has been summarized recently in book form by the present author [2]. For this reason, only a few highlights will be mentioned below, to indicate the general scope of the method and future directions of research.

Figure 3 shows a typical example of the structural information that can be derived from a rotatory-dispersion curve. It contains the dispersion curves of three isomeric cholestanones, distinguished only by the location of the carbonyl group; it will be noted that they differ significantly either in shape or in sign. A point of great importance is that these characteristic curves persist in the presence of many other substituents, such as esters, acids, alcohols, and lactones, that do not absorb in the same spectral range as the carbonyl group. Consequently we have now a tool that can frequently tell us, using less than 1 mg of recoverable substance, where a keto group, and hence the corresponding hydroxyl function, is located in an unknown steroid. At present, no other physical method will give this information.

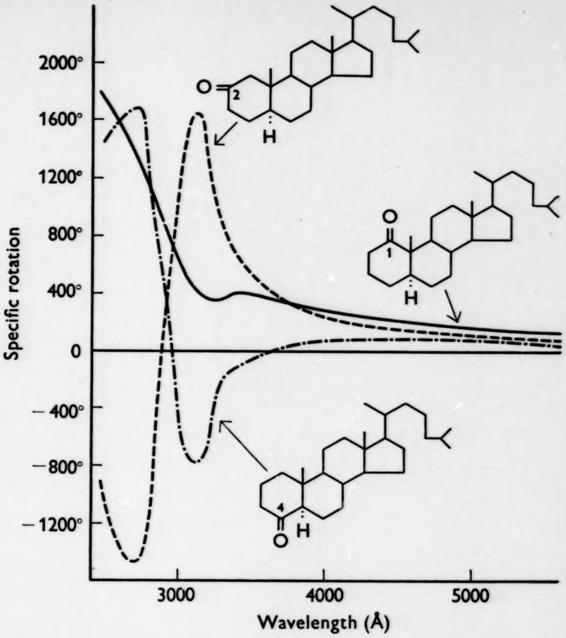


FIGURE 3 – *Optical rotatory-dispersion curves of some isomeric cholestanones.*

The differences in the rotatory-dispersion curves of figure 3 were brought about by moving the carbonyl group around the asymmetric framework and thus placing it each time in a distinct asymmetric environment; this change is reflected in the dispersion curve. Alternatively, such a change in environment can be produced by keeping the chromophore in the same place and altering the stereochemical environment. Thus, reduction of the double bond of the male sex hormone testosterone yields two isomers, 5α- and 5β-androstan-17β-ol-3-one, differing only in the nature of the A/B ring juncture. As shown in figure 4, the 5α-isomer exhibits a positive (A) and the 5β-isomer, a negative (B) Cotton effect. These characteristic rotatory-dispersion features remain unchanged, even if the substituents at C-5 or C-10, or both, in these molecules are replaced by other 'non-chromophoric' groups, such as CH_2OH, CN, or CO_2CH_3; this has proved to be very useful in the steroid field for determining the nature of the A/B ring juncture in an unknown 3-keto steroid.

When the carbonyl group is placed next to the ring juncture, equilibration between the two isomers is possible in the presence of acid or base. This is illustrated in figure 5 for the two 15-ketones of the bile-acid series, which differ only in the configuration at C-14. The C/D *cis* isomer (B), with a negative Cotton effect, is the more stable isomer, and the kinetics of the isomerization can

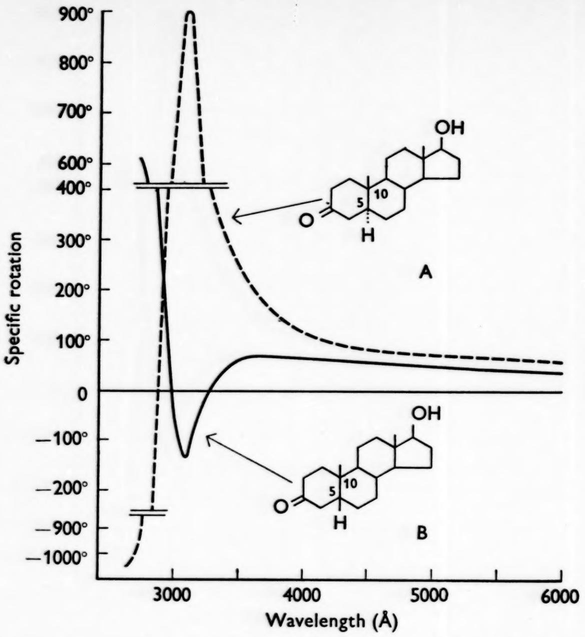


FIGURE 4 – *Optical rotatory-dispersion curves of 5α-androstan-17β-ol-3-one (dihydrotestosterone) (A) and 5β-androstan-17β-ol-3-one (B).*

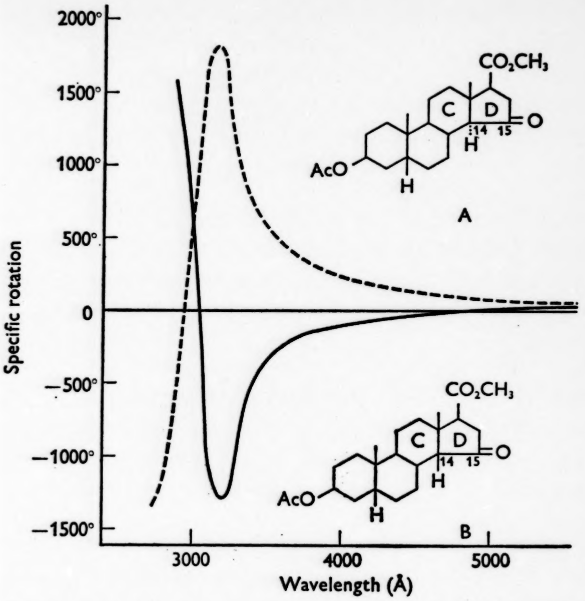


FIGURE 5 – *Optical rotatory-dispersion curves of methyl 3β-acetoxy-15-oxoetianate (A) and methyl 3β-acetoxy-15-oxo-14β-etianate (B).*

be determined very conveniently on a micro-scale by taking advantage [7] of the information contained in figure 5. It will be noted that the rotations of the two isomers are rather similar at the sodium D line in the visible, whereas in the ultraviolet near 3200 Å the difference in specific rotation amounts to over 3000°. All that is necessary, therefore, is to measure the rate of change in rotation at some convenient wavelength in the ultraviolet, and to calculate from these data the kinetics and the position of equilibrium.

Figures 3, 4, and 5 represent typical examples from the steroid series of structural, stereochemical, and analytical-kinetic applications of dispersion measurements. That these generalizations from the steroid group are, in fact, of much wider scope was demonstrated [8] by a process of dissection, by which it was found that only the immediate asymmetric bicyclic environment of the carbonyl group is responsible for the characteristic rotatory-dispersion features. This conclusion led immediately to the two most important applications of rotatory dispersion, namely the assignment of absolute configurations and the detection of conformational alterations.

APPLICATIONS IN CONFORMATIONAL ANALYSIS

The use of this tool in conformational analysis is illustrated in figure 6. *Trans*-10-methyl-2-decalone (I) represents the bicyclic analogue of a 5α-3-keto steroid (A in figure 4), and its conformation can be defined unambiguously as the all-chair form IA (figure 6). The rotatory-dispersion curve of this ketone I proved to be practically identical with that of steroid A in figure 4. A considerably more difficult problem concerns the isomeric *cis*-10-methyl-2-decalone (II in figure 6), where two all-chair conformations (IIA and IIB) have to be considered. Of these, IIA is called the steroid conformation, as steroids are frozen in this conformation by virtue of the additional ring fusions. We had already established [8] that the more distant rings in the steroid series do not play an important effect as far as the typical shape of the rotatory-dispersion curve of a given ketone is concerned; the negative Cotton effect of the steroid B in figure 4 will thus serve as a suitable reference curve for the steroid conformation of the bicyclic ketone II in conformation IIA (figure 6). It will be seen that, in the non-steroid conformation IIB, the carbonyl group finds itself in a different asymmetric environment, and its rotatory-dispersion curve would therefore be expected to be quite different from that of the steroid model. When II was synthesized [8] in optically active form and its rotatory-dispersion curve was measured, it proved to be extremely similar to that of the

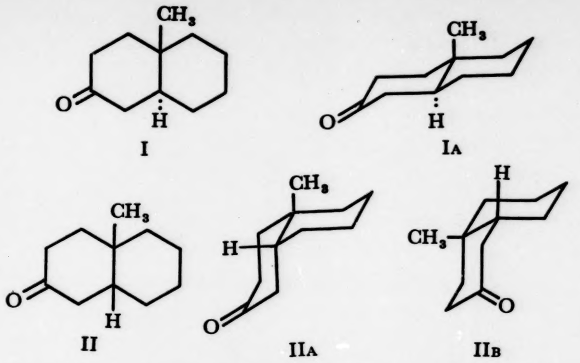


FIGURE 6 – *Conformational representations of* trans-10-*methyl-2-decalone (I) and* cis-10-*methyl-2-decalone (II).*

steroid B in figure 4. One can deduce, therefore, that the ketone II exists in the steroid conformation IIA, a conclusion which, *a priori*, was not self-evident to the stereochemist. At the present time, rotatory dispersion has been the only physical method that has shed any light on the conformation of such a decalone in solution.

THE DETERMINATION OF ABSOLUTE CONFIGURATION

The determination of absolute configuration, that is, the differentiation between mirror-image representations, has been of concern since the earliest days of stereochemistry. The only direct procedure is that of X-ray analysis [9], but so far this has been used in only very few instances. What is really needed is a quick method that permits the establishment of the absolute configuration of a given molecule by relating it to another reference compound whose absolute configuration has been settled directly or indirectly (after chemical transformation) by the X-ray method. Optical rotatory-dispersion measurements have satisfied this very important need.

In principle, the approach is quite simple [8] and is based on the same premise as was employed for examining the conformation of *cis*-10-methyl-2-decalone. We have shown that the characteristic shape of a rotatory-dispersion curve is governed by the bicyclic stereochemical and conformational environment of the chromophore—in the present instance, the carbonyl group. As additional non-chromophoric substituents do not interfere, all that is necessary is to compare the rotatory-dispersion curve of the unknown compound with that of a known reference substance possessing the same chromophore in a similar environment.

The absolute configuration of the steroids is known [10] and their rotatory-dispersion curves, therefore, can be used for such reference purposes. A typical example is reproduced in figure 7, which shows rotatory-dispersion curves of the steroid Δ^4-cholesten-3-one (I) and the sesquiterpene carissone (II). The hydroxy-isopropyl substituent at C-7 of carissone represents a non-chromophoric substituent, as does the extra methyl group at C-4. This latter point was established [8] in a separate experiment in which it was found that the rotatory-dispersion curve of Δ^4-cholesten-3-one (I) was not altered in any of its essential details by the addition of a methyl group at C-4. Consequently, the carbonyl chromophore in the steroid I of known absolute configuration and in the sesquiterpene carissone (II) must be situated in identical bicyclic environments. As their rotatory-dispersion curves are very similar in shape and sign, it is certain that the absolute configuration of carissone is that represented in formula II, rather than its mirror image. The third substance in figure 7, the tricyclic ketone III, also has the same relative stereochemical environment in rings A and B as the steroid I. Its rotatory-dispersion curve, however, is essentially of mirror-image type, from which it follows that, in terms of its absolute configuration, ketone III belongs to the antipodal series.

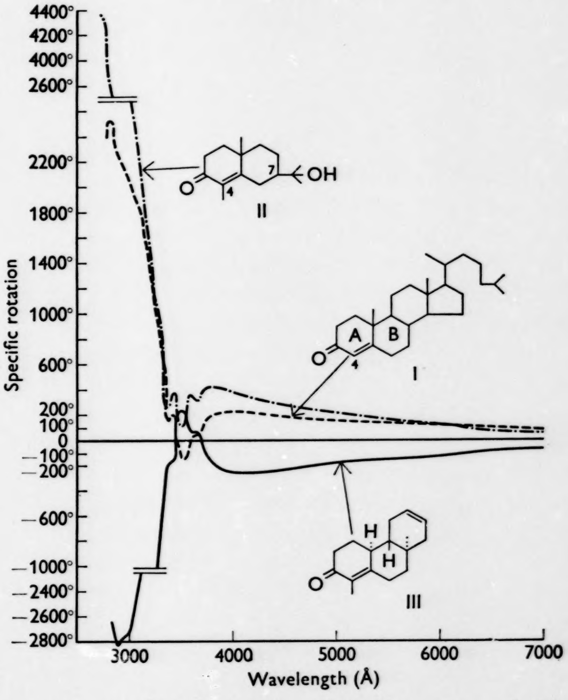
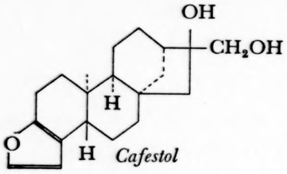
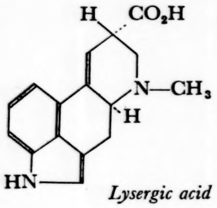
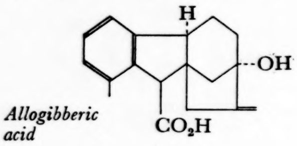


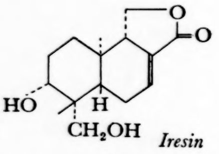
FIGURE 7 – *Optical rotatory dispersion curves of* Δ^4-*cholesten-3-one (I), carissone (II), and* (—)-1,14-*dimethyl-2-oxo-*$\Delta^{1(11),6}$-*decahydrophenanthrene (III).*

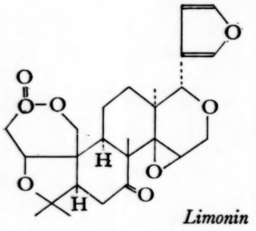
Various refinements of this approach have been introduced recently, including semi-empirical generalizations that often make it possible to predict the sign of the Cotton effect of an unknown ketone without requiring the rotatory-dispersion curve of a reference compound. These generalizations have already been reviewed elsewhere [11]. Suffice it to say that by the approach illustrated in figure 7, or by extensions of this approach [11], the absolute configurations of a wide variety of very complicated natural products have been established very quickly. A few typical examples are listed below, and it is obvious that any other classical approach would have required a tremendous effort in order to arrive at the same conclusions.


Cafestol


Lysergic acid

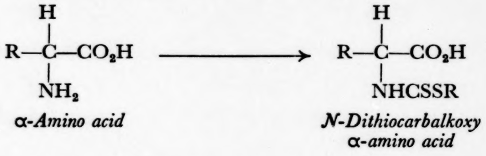

Allogibberic acid


Iresin


Limonin

Most of the discussion until now has centred on the carbonyl chromophore. Lately, our attention, as well as that of other investigators, has been directed to other chromophores, and very substantial progress has already been achieved. Recent examples are dithiocarbamates and xanthates [12], disulphides and diselenides [13], phthalimides [14], acylthioureas [15], and osmates [16]. This partial list is important not only because it illustrates the range of different chromophores that may give rise to Cotton-effect curves, but even more significantly because it includes derivatives of important non-chromophoric functional groups.

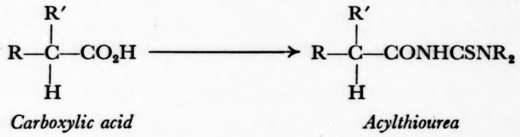
For instance, α-amino acids do not absorb in a convenient region of the ultraviolet, and they thus show only plain rotatory-dispersion curves; however, their readily prepared N-dithiocarbalkoxy derivatives (dithiocarbamates) exhibit strong Cotton effects. In fact, the sign of the Cotton effect can be related directly [12] to the absolute configuration of the asymmetric centre, members of the L-series showing a positive Cotton effect and of the D-series a negative one.



$$\mathrm{R{-}\underset{NH_2}{\overset{H}{C}}{-}CO_2H} \longrightarrow \mathrm{R{-}\underset{NHCSSR}{\overset{H}{C}}{-}CO_2H}$$

α-Amino acid *N-Dithiocarbalkoxy α-amino acid*

Similarly, carboxylic acids do not absorb above 2100 Å, but their derived acylthioureas [15] contain a low-intensity absorption maximum near 3400 Å. It is not surprising, therefore, that such derivatives should yield Cotton-effect curves, and that the sign of these Cotton effects can be used to deduce the absolute configuration of the α-asymmetric centre of the parent carboxylic acid.



$$\mathrm{R{-}\underset{H}{\overset{R'}{C}}{-}CO_2H} \longrightarrow \mathrm{R{-}\underset{H}{\overset{R'}{C}}{-}CONHCSNR_2}$$

Carboxylic acid *Acylthiourea*

Olefins are essentially transparent above 2000 Å, but they react readily with osmium tetroxide to form osmate esters. These esters, and especially

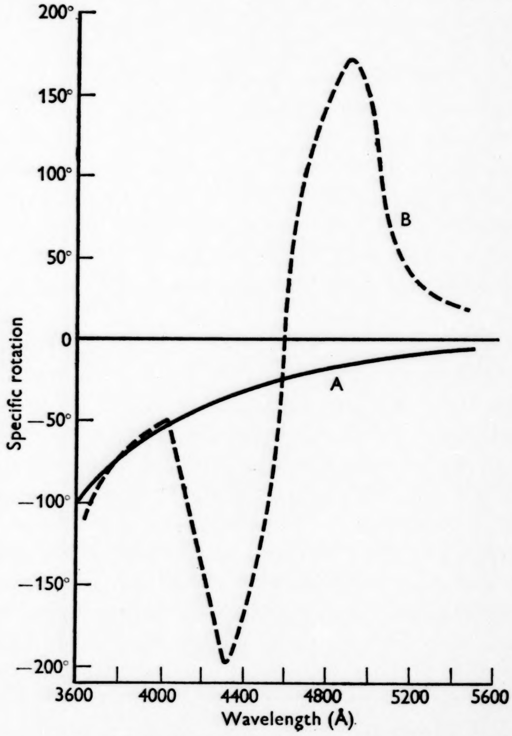


FIGURE 8 – *Optical rotatory-dispersion curve of poly-α-L-glutamic acid at pH 4·9. Curve A, polypeptide without dye; curve B, polypeptide-acriflavine complex (redrawn from Blout, E. R. and Stryer, L. [19], by permission of the authors).*

their dipyridyl complexes, have been found [16] to exhibit strong Cotton effects in a very convenient region of the spectrum. As these osmates can be prepared on a micro-scale, there is now available a means of studying certain structural and stereochemical features of olefins in terms of anomalous optical rotatory-dispersion curves.

THE CONFORMATION OF POLYPEPTIDES AND PROTEINS

One very exciting application of optical rotatory dispersion is the determination of the conformation of polypeptides and proteins. As this has been reviewed recently in great detail by E. R. Blout [17], only brief reference will be made to it here. The hydrogen-bonded α-helix structure of proteins, initially proposed by L. Pauling and R. B. Corey, is now generally accepted. One of the most effective ways of studying such a conformational feature has been by optical rotatory dispersion, even though such curves have until recently been of the plain type. Specifically, it was found that the plain curves of polypeptides in a random conformation follow a simple Drude equation (this relates optical rotatory power and wavelength outside the region of an optically active absorption-band), although this is not true of the helical conformation. The consequent mathematical generalization, largely due to W. Moffitt, has proved to be very useful. M. Goodman and collaborators [18] have demonstrated by such rotatory-dispersion measurements that the critical range where synthetic polypeptides of γ-methyl glutamic acid shift from the random to the helical structure falls between the hepta- and nona-peptide.

Two important recent observations [19, 20] indicate that anomalous optical rotatory dispersion with experimentally determined Cotton effects will play an increasingly significant role in the polypeptide and protein fields. In the first paper it was noted that while poly-α L-glutamic acid exhibits only a plain rotatory-dispersion curve (A in figure 8), its complex with the optically inactive dye acriflavine shows a Cotton effect (B in figure 8) centred around the 4570 Å absorption band of the dye. Such Cotton effects are observed with the dye-polypeptide complex only when the polypeptide is in the helical rather than the random conformation. The complexes thus offer a strikingly simple approach to the study of helical conformations, as well as to the determination of the 'handedness' of the helix. Of even more general applicability is the observation that the Cotton effects of polypeptides and proteins may be determined directly in the 2100–2300 Å region with the new automatically recording Rudolph spectropolarimeter.

REFERENCES

[1] Biot, J. B. *Mém. Inst.*, **1**, 1, 1812; *Mém. prés. Acad. Sci., Paris*, **2**, 41, 1817.

[2] Djerassi, C. 'Optical Rotatory Dispersion: Applications to Organic Chemistry'. McGraw-Hill, New York. 1960.

[3] Lowry, T. H. 'Optical Rotatory Power'. Longmans, Green & Co., London. 1935.
Levene, P. A. and Rothen, A., in H. Gilman, 'Organic Chemistry', Vol. II, Chapter 21. John Wiley & Sons, New York. 1938.
Kuhn, W. *Ann. Rev. Phys. Chem.*, **9**, 417, 1958.

[4] See chapter 2 in [2] and Djerassi, C. and Klyne, W. *Proc. chem. Soc., Lond.*, 55, 1957.

[5] See chapter 16 in [2].
Jones, P. M. and Klyne, W. *J. chem. Soc.*, 871, 1960.
Sjöberg, B. *Ark. Kemi*, **15**, 451, 1960.

[6] Djerassi, C. *et al. J. Amer. chem. Soc.*, **77**, 4354, 4359, 4364, 1955; *ibid.*, **78**, 440, 3163, 3761, 1956.

[7] Allinger, N. L., Hermann, R. B. and Djerassi, C. *J. org. Chem.*, **25**, 922, 1960.

[8] Djerassi, C., Riniker, R., and Riniker, B. *J. Amer. chem. Soc.*, **78**, 6362, 1956.
Djerassi, C. and Marshall, D. *Ibid.*, **80**, 3986, 1958.

[9] Bijvoet, J. M. *Endeavour*, **14**, 71, 1955.

[10] Fieser, L. F. and Fieser, M. 'Steroids'. Reinhold Publishing Corporation, New York. 1959.

[11] See chapters 9, 10, and 13 in [2].

[12] Sjöberg, B., Fredga, A., and Djerassi, C. *J. Amer. chem. Soc.*, **81**, 5002, 1959.

[13] Djerassi, C., Fredga, A., and Sjöberg, B. *Acta chem. Scand.* (In press.)

[14] Brewster, J. H. and Osman, S. F. *J. Amer. chem. Soc.*, **82**, 5754, 1960.

[15] Djerassi, C. and Undheim, K. *Ibid.*, **82**, 5755, 1960.

[16] Bunnenberg, E. and Djerassi, C. *J. Amer. chem. Soc.*, **82**, 5953, 1960.

[17] Blout, E. R. Chapter 17 in [2].

[18] Goodman, M., Schmitt, E. E., and Yphantis, D. *J. Amer. chem. Soc.*, **82**, 3483, 1960.

[19] Blout, E. R. and Stryer, L. *Proc. Nat. Acad. Sci., Wash.*, **45**, 1591, 1959.

[20] Simmons, N. S. and Blout, E. R. *Biophys. J.*, **1**, 55, 1960.
Simmons, N. S., Cohen, C., Szent-Györgyi, A. G., Wetlaufer, D. B., and Blout, E. R. (Unpublished observations.)

The nucleolus of the cell nucleus

J. L. SIRLIN

The nucleolus, an organelle of the cell nucleus, was first described by Fontana nearly two centuries ago, but its nature remained obscure until recently. Its precise function in the cell has still not been completely elucidated, but it is known to be involved in protein synthesis. This article discusses current theories of the biochemical function of the nucleolus.

'*On observe un corps oviforme, ayant une tache dans son milieu.*' Fontana used these words in 1781 [1] in describing cells in the slime of an eel; it is the first mention of the nucleolus; the nucleus itself, Fontana's oviform body, had been first described by Leeuwenhoek in 1702. The nucleolus is present in most cell nuclei [2]. In general it is a roundish body, numbering one or a few per nucleus, and denser than the remainder of the chromatin, in which it is embedded, because of its high protein content. The subtlety of the function of the nucleolus is such that it was not described in general chemical terms until the 1940s. This was largely a result of the development, by the schools of T. O. Caspersson in Sweden and J. Brachet in Belgium, of knowledge of the cell chemistry of ribonucleic acid (RNA) and protein. The precise chemistry of the function is still not clear.

Apart from RNA, which is its most characteristic constituent, the nucleolus contains lipids, carbohydrates, and mineral substances. Strictly speaking, it lacks deoxyribonucleic acid (DNA), the most characteristic constituent of chromosomes. However, it is surrounded, in most cytological preparations, by condensed chromatin (the distended chromosomes) forming the nucleolus-associated chromatin. Because this material is considered by some authors to be part of the nucleolus, the latter is often said to contain DNA. Sometimes the nucleolus-associated chromatin is of the type known as heterochromatin.

Inside the nucleus, the nucleolus is subject to movements. These may be either passive, due to the gel-like consistency of chromatin, or active, such as emission of prolongations, amoeboid movements, fusion with other nucleoli and the nuclear membrane, and even fragmentation; the active are the more important. This plasticity of the nucleolus may lead to very bizarre shapes, as in certain nurse cells and in ciliate protozoa, in which it departs considerably from the usual roundish appearance. The movements of the nucleolus are usually more pronounced during periods of active cell synthesis.

Because nucleoli can fuse with one another, their number is normally variable and cannot be used as a reliable measure of nucleolar activity. In most cells the general trend is for the number of nucleoli to decrease as the cell ages, except for an occasional budding that increases the number of nucleoli; in a few transient types of cell, such as oocytes and nurse cells, the number of nucleoli may increase regularly. The maximum number is fixed. It is determined by the nucleolar organizers; these are differentiated segments of certain chromosomes and therefore occur in fixed numbers under normal conditions [3]. The maximum nucleolar number is therefore an effective index to the degree of ploidy, that is, the number of sets of chromosomes. The maximum number will be found in young cells, before nucleolar fusions have taken place [4].

The exact chemical function of the nucleolus is still not fully understood. However, the fact that the nucleolus and the cell itself are active at the same time was quickly recognized, and it soon became clear that the nucleolus was involved in the cell's functions; eventually it was realized that it was particularly concerned with protein synthesis. During periods of intense protein-production the nucleolus tends to become single and large, rather than multiple and small. This is typical, for example, of gland cells secreting enzyme and mucus protein, of neutrons elaborating Nissl substance (a ribonucleoprotein), and of embryonic and regenerating cells that show a marked production of protein. Again, nucleolar development is correlated with the nutritional level of the cell. However, nucleoli in tumour cells do not show any special characteristics, and although there is a loose correlation between nucleolar appearance and malignancy it is not specific enough to help diagnosis. Nevertheless, some still believe that cancerous growth is initiated by specific stimulation of the nucleolus.

Cells that normally do not actively synthesize protein, such as mature leucocytes and muscle cells, have inconspicuous nucleoli. Similarly, the nucleolus normally disappears during the phase in the cell cycle when the cells are not synthesizing, that is, during cell division (the spindle protein is synthesized before cell division), and cleaving embryonic cells, that are increasing in number without making any protein, have, normally, no nucleoli. Nucleoli reappear in the embryo only at gastrulation, when maternal yolk reserves are exhausted and protein synthesis is resumed. The embryos of mammals, molluscs, and annelids, which have well-developed nucleoli from the earliest stages, are exceptions to this rule. It is not known whether this is because these nucleoli have special functions, although this is suggested by the specialized mode of embryonic development in these phyla, placental in mammals and spiral in the others. There is general, but not complete, correlation between brisk protein-synthesis in cells and their having vigorous nucleoli. Exceptions include protein-deprived liver cells and certain degenerating tumour cells. In the liver cells, the large nucleoli may represent an abortive compensatory development of the protein-synthesis system, or they may indicate, as has been suggested for neurons, a preparation for later demands on synthesis.

The fact that nucleoli are often damaged by agents that depress the cell's metabolic activity suggests that the nucleolus is involved in that activity, as does the fact that pathological conditions of cells are frequently accompanied by nucleolar changes.

It has frequently been claimed that oocytes and cancer cells produce nucleolar substance that is visible under the microscope, but the significance, and even the general occurrence of this is still doubtful, especially as it is rarely observed with the electron microscope. This instrument does, however, clearly show that there are frequently nucleoli on one side of the nuclear membrane, close to endoplasmic reticulum (an important synthetic site in cytoplasm), the Golgi complex, or mitochondria, on the other (figures 3 and 4). This certainly suggests some transfer of nucleolar material, even if only of molecular size. The association of the nucleolus with certain differentiated regions of cytoplasm is sometimes mediated by chromatin condensed around the nucleolus [5]; this was known for many years, and led to hypotheses of nucleolar function that will be discussed later.

THE APPEARANCE OF THE NUCLEOLUS

When treated with dyes, the nucleolus shows certain inhomogeneities, such as vacuoles, inclusions, and filaments. Sometimes these variations correspond to physical artefacts rather than to permanent structures, but at other times they have a real chemical basis and indicate the degree of nucleolar activity. For instance, the formation of large vacuoles is characteristic of highly secretory states, and, sometimes, of degenerative states.

The electron microscope shows the nucleolus as a fine filamentous matrix, presumably proteinaceous, in which numerous denser particles, measuring about 150 Å in diameter, are embedded (figure 4). These are similar to those found elsewhere in the nucleus and in the cytoplasm, and are assumed to consist of ribonucleoprotein. The core of the nucleolus appears denser than the more peripheral region, especially after fixation with osmium; this density may be due either to a denser matrix or to closer packing of particles. This core corresponds to the 'nucleonema' of C. Estable and R. J. Sotelo; the periphery to their 'pars amorpha' [6]. The chromosomal nucleolus organizer often remains embedded in the nucleolus that has grown from it. The appearance of the ultrastructure of the nucleolus depends markedly, as would be expected at this morphological level, on the procedures, such as fixation and embedding, used to study it, but much of the variation undoubtedly reflects physiological changes.

Differentiation of the nucleolus parallels, in certain cells, differentiation of the cell itself. Such cells include oocytes, where nucleoli can be very well developed, and certain special cells of the salivary glands of insect larvae. It has, in fact, been suggested [7] that nucleoli are, on the whole, more concerned with cell differentiation than are chromosomes, and, what is almost equivalent, that the nucleolus is the seat of general cell-heredity (cytoplasmic heredity), while the genes are responsible for the inheritance of more particular traits [8]. This links up with A. M. Dalcq's idea of 'ontomutations', which, by dramatically altering the heredity, would be the source of new major developmental traits [9]. In fact Dalcq has dwelt on the possible relationship between the presence of nucleoli in the early mammalian embryo and its placental development. Some embryos, however, are known to differentiate to some extent, although not to full term, even when they entirely lack organized nucleoli: the implications of this are discussed later.

The nucleolus may best be considered as a differentiated locus of a particular chromosome. This is particularly so in specialized chromosomes, such as the polytene chromosomes of flies; the situation is probably less simple in ordinary chromosomes. There are other chromosomal organelles, such as puffs and Balbiani rings, that, although different in some respects from nucleoli, have several similar physiological and morphological characteristics. Again, there is normally no organized nucleolus in the salivary nuclei of chironomid flies, but, instead, a nucleolar substance spread rather widely on the chromosomes. It is, in fact, possible to speak generally of a nucleolar substance of which the nucleolus is an organized form.

It is not known whether all nucleoli, which presumably have largely similar functions, are similar in origin. This question arises in the case of nucleoli in cells formed during the early division of mammalian eggs, which nucleoli may originate from both chromosomes and other sources. It also arises, especially, in the case of the nucleoli of amphibian oocytes, where some nucleoli are apparently formed as their somatic counterparts in chromosomes, while others are definitely not formed there [10]. Similarity of origin in these cases would be possible only if there were a general nucleolar substance; the evidence appears to suggest that there is.

It is still questionable whether bacteria and and blue algae have nucleoli, but protozoa certainly have them. The organisms without nucleoli are highly successful, by biological criteria; this raises the question, which will be considered later, of the relevance of nucleolar function in the more highly evolved cells.

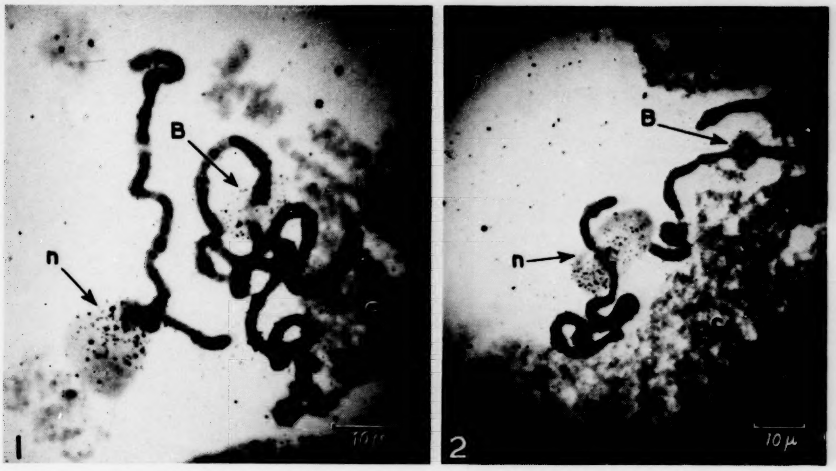
The nucleolus disappears at the end of prophase, before division of the cell, and reappears at telophase after division. Sometimes the nucleolus persists through cell division, being discarded into the cytoplasm, but new nucleoli are formed even in these cases. These persistent nucleoli are more typical of plants than animals; they are associated with certain other morphological characteristics, and so have proved to have a limited taxonomic value. There are a number of possible explanations of the disappearance of the nucleolus. It may occur at the same time as a change in nuclear conditions that is necessary if nucleolar integrity is to be maintained, possibly by making the nucleolus accessible to enzymes that are restricted to cytoplasm during interphase. Another possibility is that nucleolar-chromosomal interactions, for example, demand periodic renewal or reformation of the nucleolus; a third is that partition of nucleoli between the two daughter cells during mitosis proves embarrassing to the cell, which resorts instead to forming new nucleoli. That the nucleolus disappears so as to release nucleolar substance seems unlikely, as new nucleoli are formed even when old ones persist.

The way in which the nucleolus forms anew at telophase offers some insight into the nature, and therefore the significance, of the nucleolar substance. In nuclei with polytene chromosomes, the nucleolus is the product of the chromosomal locus on which it is borne; in ordinary nuclei the nucleolus forms at the organizer, but uses, partly at least, materials formed elsewhere in the nucleus. Microscopy and electron microscopy show, fairly convincingly, 'prenucleolar bodies' in such nuclei, forming amidst the chromosomes and gathering at the organizer to re-form the nucleolus; exceptionally, the organizers are missing and the bodies remain ungathered. The driving force behind the coalescence of the bodies remains obscure, and there is evidence for the control of nucleolus formation by widespread genetic loci as well as by the organizer. There are also several lines of evidence for a competition for nucleolar precursors between different organizers co-existing in a nucleus or, alternatively, for a mutual inhibition between sometimes quite numerous potential organizers in the same nucleus. Thus, when organizers are dislodged into nuclear 'pockets' they form smaller nucleoli, as if they were cut off from precursors by the more centrally located organizers; these form nucleoli of normal size. Again, after irradiation, the lagging chromosomes and the micronuclei derived from them produce a total amount of nucleoli that is greater than normal. The balance of these intra-nuclear interactions favours, in normal circumstances, the predominant organizer, which will form the nucleolus. Swarming of nucleoli, as it occurs in oocytes, pronuclei, and protozoa is thus explainable by a temporary relaxation of competition or inhibition.

THE FUNCTION OF THE NUCLEOLUS

The nucleolus was initially considered to form the nucleus, and that in turn to form the cell [11]. Until recently, it was also thought to give rise to yolk and melanin, and to all the organelles of *Paramecium* [12]. Various other functions proposed have been that the nucleolus is a source of nutritional reserves for the cell in general or for the nucleus in particular; an energy reservoir for the nucleus; a contribution from cytoplasm to nurture the nucleus; a reservoir for materials produced at a limited rate by the chromosomes; and a store for waste from nuclear and cytoplasmic metabolism.

The 'matrix theory' proposes that when the



FIGURES 1 *and* 2 – *Autoradiograms showing the beginning of incorporation of tritiated uridine, a precursor of RNA, in the nucleolus* (*n*) *and Balbiani ring* (*B*) *of salivary gland nuclei of a chironomid. In the nucleolus the incorporation begins close to, but not in, the chromosomal nucleolus organizer at the centre of the nucleolus. Chromosomes and cytoplasm* (*c*) *have not yet begun to incorporate tracer.* (*Reproduced by permission of Butterworths Scientific Publications, London.*)

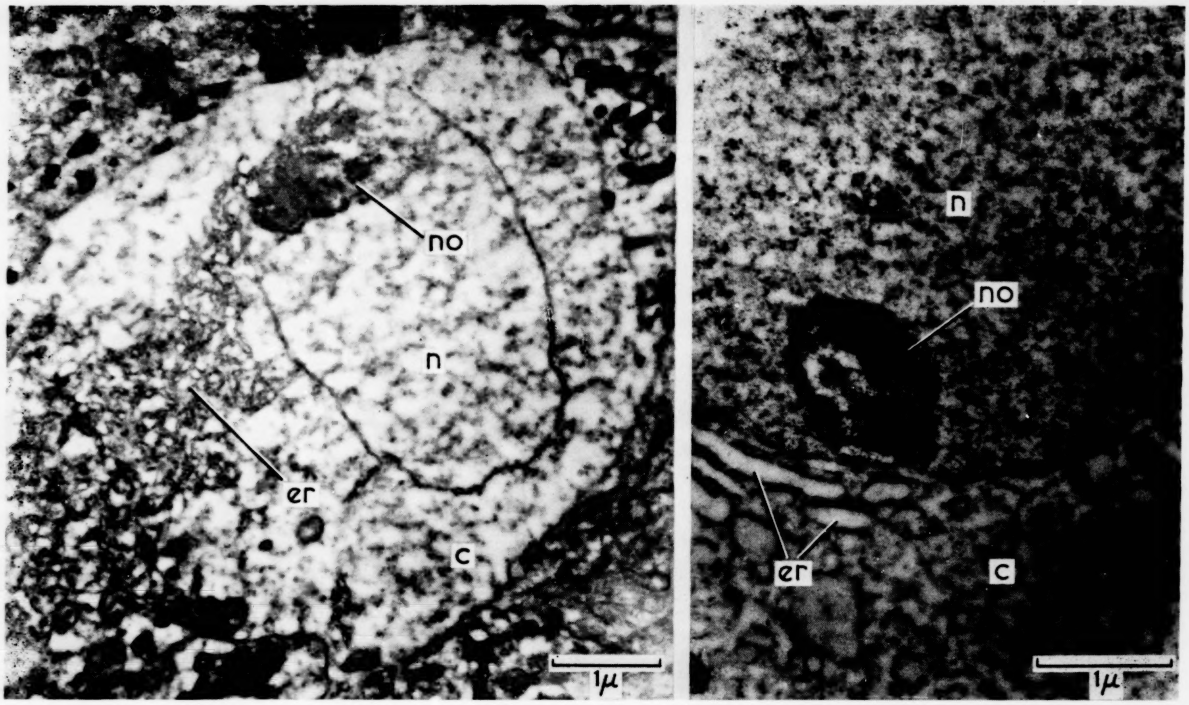
nucleolus disappears at division, nucleolar materials are deposited as a matrix on the chromosomes, and re-enter the new nucleolus formed at telophase. C. D. Darlington [13] speaks of the nucleolus and heterochromatin as alternative sources of nucleic acid for the chromosomes. Certain appearances of the nucleolar apparatus in nerve and gland cells, when it comes into contact with the nuclear membranes through the nucleolus-associated chromatin, suggest that nucleolar substance is discharged into cytoplasm.

Two stimulating modern theories have been put forward by Caspersson and his associates [5], and by Estable and Sotelo [6]. Caspersson considered that the nucleolus-associated chromatin, when stimulated, would produce basic proteins that would gather at the nucleolus and then pass to the cytoplasm, where the synthesis of specific cell-proteins would take place. RNA would be operative in both the nucleolar apparatus and the cytoplasm, the apparatus directing the cell system for protein synthesis. According to this theory, the mechanism gets out of control in cancer cells, and consequently operates with excessive intensity. If the roles attributed in this theory to basic proteins and to nucleolus-associated chromatin are extended to include RNA and chromatin as a whole, the main tenets of this theory remain essentially valid today. Caspersson has recently considered the passage of nuclear RNA to cytoplasm.

Estable and Sotelo regarded the nucleolus as formed of two components of different significance. One, the nucleonema, is deposited on the chromosomes during cell division, and enters the new nucleolus that forms after division; the nucleonema has genetic continuity (*omnis nucleonema e nucleonema*). The second component, the pars amorpha, is a plastic one that disappears at mitosis and forms again at telophase; the pars amorpha would be the metabolic component for each functional phase of the nucleolus during interphase. This conceptually attractive theory, which retains elements of the matrix theory, is more difficult to evaluate from the available evidence than is Caspersson's. There is not enough convincing evidence that the nucleonema is deposited on the dividing chromosomes; such evidence would, of course, be difficult to obtain. What is probably a more serious criticism is that of J. A. Serra [14]. He argues cogently, on the basis of his own observations, for an epigenetic, rather than a genetic, role for the nucleonema.

The discussion that follows deals with the chemical processes that might be performed by the nucleolus; a systematic theory has not been attempted and the ideas are offered to suggest further work. The discussion is based on the assumption that the nucleolus plays a general part in the cells of higher organisms that is indispensable in the long run to cells that produce protein, but essentially auxiliary.

The assumption that the function is general results from the identity of the nucleolar locus in the chromosome (the organizer) in all nuclei of an organism and from the functional equivalence of nucleoli as shown in interspecific hybrid crosses. That it is indispensable for protein synthesis is suggested by the facts that the nucleolus is generally present when the cell is synthesizing protein and that it is an entity, without a membrane, embedded in the genetic code (the cell nucleus); by the weighty circumstantial evidence for its metabolic activity; and by its turnover of RNA, which is known to participate in protein synthesis. Finally, the following facts suggest that the role of the nucleolus is essentially auxiliary. Firstly, embryonic differentiation, which implies synthesis of



FIGURES 3 *and* 4. – *Electron-micrographs of cone cell of larval insect eye (figure 3) and of notochord cell of amphibian embryo (figure 4), showing the nucleus (n) and the nucleolus (no). The nucleolus appears to be associated with developing endoplasmic reticulum (er) in the cytoplasm (c). (By courtesy of C. H. Waddington and M. M. Perry.)*

specific protein, proceeds to some extent in the absence of organized nucleoli and of organizers. Secondly, cells that do not synthesize protein, and cleaving embryos—both on a 'maintenance metabolism' as it were—have no conspicuous nucleoli, although they still require, for example, maintenance of enzymes. Thirdly, lower organisms, some cells in sciarid larvae, and rapidly dividing tumour cells synthesize proteins, although they seem to have no organized nucleoli. The role of the nucleolus may be the result of its having either one function in cell metabolism or a complex set of functions that acts, for instance, as a rate-limiting system. The nucleolus may also, in certain cases, have specialized functions; these are not considered here. They may result from secondary functional adaptations and need not exclude a more general nucleolar function, which will be discussed in relation to enzymes, RNA, and protein.

ENZYMES

Nucleoside phosphorylase and DPN synthetase, two mainly nuclear enzymes that presumably participate in the synthesis of coenzyme, are localized in the nucleolus [15], and it is therefore possible that the nucleolus is a source of coenzyme for the cell. It has been suggested that when nucleoli and mitochondria occasionally come closer together across the nuclear membrane, the nucleoli provide the mitochondria with coenzyme. Dalcq has proposed that the nucleolus produces adenosine triphosphatase in embryos.

RIBONUCLEIC ACID

The participation of the nucleolus in the RNA turnover of the cell is likely to be its most important function, consisting in making RNA either as the end-product or as an intermediate in the synthesis of some protein. This is one aspect of the wider question of the provision or control by the nucleus of the RNA needed by the cytoplasm for protein synthesis. The dependence of cytoplasmic RNA on that in the nucleus is supported by ample evidence, and the intensity of nucleolar RNA metabolism is in general parallel to that of nuclear RNA. Recent experiments on irradiation of the nucleolus have demonstrated that part of the cytoplasmic RNA is dependent on that of the nucleolus, and to a lesser extent, on chromosomal RNA [16]. This was to be expected from the fact that the nucleolus contains a fair proportion of the total RNA in the nucleus but rarely the whole of it. It is important to note that these experiments also show that a fraction of the cytoplasmic RNA is not dependent on that in the nucleus or nucleolus. This fraction may be synthesized *in situ* in cytoplasm instead of in the nucleus, where the nuclear- or nucleolar-dependent cytoplasmic RNA fractions are supposed to be synthesized. It has recently been shown that, in spider oocytes, the nucleotide composition of nucleolar and cytoplasmic RNA is identical, and different from that of the remaining nuclear RNA; this, again, suggests some correspondence between the two former RNAs [17]. The genetic implications of the nuclear origin of cytoplasmic RNA will become clear later.

The local synthesis during interphase of at least a part of the nucleolar RNA is safely inferred from the existence in the nucleolus of all the precursors and enzymes for RNA synthesis, and from biochemical and autoradiographic studies. Generally, the more differentiated the cell is, the greater the rate of synthesis. During reconstitution of the nucleolus at telophase some of the nucleolar RNA is apparently contributed by the chromosomes in the form of the prenucleolar bodies already described. One possibility to be seriously considered is that of a continuous contribution of chromosomal RNA to that in the nucleolus even during interphase. Nucleolar function during interphase and nucleolar reformation during telophase would then not be different processes but two aspects of the same process. The impression gained from the evidence as a whole is that the turnover of RNA is a property of the nucleolar substance rather than, for instance, of the organizer, irrespective of whether it has coalesced as an organized nucleolus or not. However, the most active site of RNA turnover for different nucleoli varies according to their organization. It is central in the chironomid salivary nucleolus, which remains attached to the central organizer (figures 1 and 2), and it is spread over the surface of the *Drosophila* nucleolus, which is detached from the organizer.

Nucleolar RNA may be related to RNAs with different biochemical functions, transfer, ribosomal, and messenger RNA. Whatever the function of nucleolar RNA may be, the nucleolus could serve *inter alia* as a store for it. This is suggested by its mode of formation, as well as by instances of nucleoli becoming surcharged with RNA as the result of altered physiological conditions. For more details the readers is referred to a review article by the author [18].

Transfer RNA exists in a low-molecular form in the cell sap; its function is to accept the amino acids presented by specific amino-acid-activating ensymes, and to deposit them in ribosomal RNA, where they are synthesized into protein. It is believed that the origins of transfer RNA may be in the break-down of ribosomal RNA, or synthesis in the cell sap or the nucleus.

Synthesis of transfer RNA was found in the two types of nucleolus that have been investigated, namely in the sea-urchin oocyte [19] and in the chironomid salivary cell [18]; it is likely that it occurs in most other nucleoli. Transfer RNA was characterized by biochemical criteria in the oocyte and by the incorporation of radioactive pseudouridine, a fairly specific precursor, in the salivary cell. It was shown that in the latter the nucleolus is an important source of transfer RNA in the cell.

The synthesis of transfer RNA indicates a definite function for the nucleolus in protein synthesis in the cell, either in the supply of this RNA as such, in its use in the nucleolar synthesis of a protein, or for storage of activated amino acid. Transfer RNA does not convey genetic information, but has instead an auxiliary function in protein synthesis that corresponds well with the overall auxiliary function in the cell contemplated for the nucleolus itself. It is, therefore, possible that the nucleolus evolved in adaptation to the function of transfer RNA in protein synthesis. This function could be a primordial one, in an evolutionary sense, or evolved from systems lacking transfer RNA but having alternative pathways, the two systems coexisting today in the same cell. Transfer RNA occurs in lower organisms that do not possess organized nuclei and nucleoli: that these organisms never develop multi-cellular complexity may prove significant.

The high-molecular RNA of ribosomes exists in the ubiquitous ribonucleoprotein particles of the cell that are concerned with protein synthesis. This serves as a template for the conversion of amino acids into protein, the acids being presented by transfer RNA. For this task ribosomal RNA should contain some information for the proper sequential alignment of amino acids. This is in principle coded in the genes, and it is thought that ribosomal RNA is synthesized in the nucleus under their control, and is later conveyed to the cytoplasmic sites of synthesis, where it fulfills its role of template. Since the nucleolus itself represents a gene, ribosomal RNA could also be directly synthesized there. Messenger RNA is assumed to programme the ribosomes, and therefore also to contain some sequence information; it would be synthesized under gene control.

Evaluation of any ultimate nucleolar contribution in ribosomal RNA remains speculative, because of the need for means of chemical characterization *in situ*. The synthesis of transfer RNA in the nucleolus, discussed in the previous section, does not exclude that of ribosomal RNA, as all the four major ribonucleosides that occur in the nucleolus are common to both types of RNA. The electron microscope shows that the ribosomal particles present in the nucleolus are similar to those in the nucleus and cytoplasm. It might be that these particles are synthesized in the nucleolus and are destined to go to the cytoplasm. They could either be completely synthesized there or only the protein moiety; in the latter event, their RNA would presumably be synthesized elsewhere in the nucleus. The biochemical evidence favours a nuclear origin for the complete particle. Alternatively, these particles could be the nucleolar site for synthesis of protein and not be destined for export to cytoplasm. Coexistence of two particle populations in the nucleolus, one being part of the local machinery of synthesis and the other the product of this synthesis, is also possible. Furthermore, it is conceivable that a single particle could display synthetic activity first in the nucleolus and later in the cytoplasm.

Currently, the nucleolus is regarded as in some way assisting in the replication of genic RNA, which otherwise would not be produced in sufficient amounts for the cytoplasmic sites of synthesis, particularly at times of very active protein-synthesis, such as during secretion or differentiation. Genic RNA is copied using master templates that are deposited in the nucleolus. During reconstitution of the nucleolus at telophase, this could be achieved through the prenucleolar bodies. This would imply a flux of nuclear RNA during interphase towards the nucleolus, but the evidence of this from autoradiography is inconclusive. If the ribosomal and messenger RNAs are gene-specific, as assumed, it follows that the nucleolus should incorporate master templates from as many genes as require nucleolar collaboration for the replication of these RNAs. It has also been suggested that nucleolar RNA could bind protein for storage, and *vice versa*.

Three types of template RNA are therefore needed in the nucleolus, to fulfil an auxiliary role in the system of protein synthesis:

1. A single template (or at most a few templates) for production of a general nucleolar RNA, transfer or ribosomal RNA. Since the nucleolus itself represents one gene (or at most a few), this template can be nucleolar RNA proper and not RNA from the chromosomes. There is, in fact, evidence that in polytene chromosomes nucleolar RNA is the product of the nucleolus itself.

2. A template for production of a general nucleolar end-product, such as ribosomal protein. This protein can, as in the case of transfer or ribosomal RNA, be produced by one or a few templates consisting of nucleolar RNA proper; it is considered later.

3. Template RNAs of chromosomal origin gathered at the nucleolus for replication of ribosomal or messenger RNA. Even if some genes require the assistance of the nucleolus to produce these RNAs, it is conceivable that not all will.

The chromosomes most likely to be self-sufficient are the multiple-stranded polytene chromosomes of diptera, and there is evidence for widespread liberation of RNA by them. (If all genes were self-sufficient, the template RNA of type 3 would be unnecessary.) Support for these views comes from the fact that polytene chromosomes of sciarid diptera normally have no nucleoli but have instead a widespread nucleolar substance, and that certain mammalian cells in tissue culture, having ordinary chromosomes, but with smaller nucleoli, also have a widespread nucleolar substance. A fundamental difference between polytene and ordinary chromosomes could thus result from the degree to which the two types require the collaboration of the nucleolus for the replication of template RNAs.

PROTEINS

Although nucleoli of both growing and mature cells possess all the prerequisites for protein synthesis, the turnover of amino acid occurs only in growing nucleoli. By contrast, RNA turns over continuously throughout the life of the nucleolus. This limited turnover of amino acid can be accepted, on various grounds, as reflecting net protein synthesis and not merely amino acid activation or exchange. Some of this synthesis would be of growing nucleolar protein itself and some of it could perhaps be of exportable protein, particularly since it has recently been found that the nucleolus utilizes, at least partly, pre-existing protein for its growth. It is clear that the fully mature nucleolus can exert control over cell proteins only through proteins previously synthesized during its growing period, or through the agency of nucleolar

RNA that continues to turn over during nucleolar maturity.

Three classes of exportable nucleolar protein will be considered. The template mechanism for producing these proteins would be as postulated for RNA in the previous section.

Several lines of evidence suggest [18] the possibility of a nucleolar production of ribosomal protein. This could combine with the ribosomal RNA to produce the finished ribosomal particle *in situ*. This is, in fact, one of the simplest overall functions that can be proposed for the nucleolus, having proper regard to its auxiliary role in the cell system for protein synthesis, which would account for all that is known of nucleolar turnover of amino acid and RNA. It implies that ribosomal particles are produced by the nucleolus only while it grows.

Several other proteins can be mentioned here. First, in some oocytes there is in both nucleoli and yolk a common antigen, which could have a nucleolar origin; it has long been supposed that the nucleolus plays a part in yolk formation. In *Acetabularia*, nucleolar proteins have been implicated in the production of morphogenetically active proteins in cytoplasm. Both these proteins could in fact be ribosomal.

Nucleolar emission of protein, sometimes as 'matrix', from the nucleus is common, if not general. It was suggested that antibody and hormone proteins are accumulated in nucleoli of the cells in which these substances are metabolized. It has also been claimed that haemoglobin is synthesized in the erythroblast by the nucleolus [20].

Several similarities between nucleoli and spindles suggest that the first might be involved in the formation of the second. Firstly, both appear to contain only one protein, as well as having other gross similarities of chemical composition. Secondly, there is an apparent association of the nucleolus with spindle function [18]. Thirdly, formation of the spindle is rapidly inhibited by irradiation of any single nucleolus in a nucleus. Fourthly, spindle protein is synthesized before spindle formation, that is before the nucleolus has disappeared at cell division and perhaps while it is still turning over amino acid. The relatively large amount of spindle protein in the cell shows, however, that not all spindle protein is likely to be directly dependent on nucleolar synthesis. An alternative proposal is that the nucleolus mediates the release of agents that are structurally active with respect to the spindle. The most obvious argument against the proposition that the only function of the nucleolus is connected with spindle formation is its manifest activity in cells that secrete but do not divide.

REFERENCES

[1] HUGHES, A. 'The Mitotic Cycle'. Butterworths Scientific Publications, London. 1952.
[2] MONTGOMERY, T. H. *J. Morphol.*, **15**, 265, 1898.
[3] HEITZ, E. *Planta*, **12**, 775, 1931.
[4] SCHREIBER, G. *Z. mikr.-anat. Forsch.*, **66**, 301, 1960.
[5] CASPERSSON, T. O. 'Cell Growth and Cell Function'. W. W. Norton, New York. 1950.
[6] ESTABLE, C. and SOTELO, R. J. *Inst. Inv. Cien. Biol. Publ.*, **1**, 105, 1951.
[7] WILLMER, E. N. 'Cytology and Evolution'. Academic Press, New York. 1960.
[8] BERNHARD, W. and OBERLING, C. 'Problèmes d'ultrastructure et de fonctions nucléaires' (edited by Thomas, J. A.), p. 279. Masson et Cie., Paris. 1955.
[9] DALCQ, A. M. *Bull. Soc. zool. Fr.*, **79**, 240, 1954.
[10] CALLAN, H. G. and LLOYD, L. *Phil. Trans., B*, **243**, 135, 1960.
[11] SCHLEIDEN, M. J. 'Contributions to Phytogenesis' (appendix to SCHWANN, T. 'Microscopical Researches into the Accordance in the Structure and Growth of Animals and Plants'). Sydenham Society, London. 1847.
[12] EHRET, C. F. 'Symposium on Information Theory in Biology' (edited by Yockey, H. P.), p. 218. Pergamon Press, New York. 1958.
[13] DARLINGTON, C. D. *Symp. Soc. exp. Biol.*, **1**, 252, 1947.
[14] SERRA, J. A. *Rev. Port. zool. biol. gen.*, **2**, 117, 1959.
[15] BALTUS, E. *Biochim. Biophys. Acta*, **15**, 263, 1954.
[16] PERRY, R. P. *Exp. Cell. Res.*, **20**, 216, 1960.
[17] EDSTRÖM, J. E. *J. Biophys. Biochem. Cytol.*, **8**, 47, 1960.
[18] SIRLIN, J. L. *Progr. Biophys. biophys. Chem.*, **12**. (In press.)
[19] VINCENT, W. S. and BALTUS, E. *Biol. Bull.*, **119**, 299, 1960.
[20] O'BRIEN, B. R. A. *Exp. Cell. Res.*, **21**, 226, 1960.

Reactive dyestuffs for cellulose

I. D. RATTEE

Dyeing is one of the most ancient of technologies, yet until little more than a hundred years ago the preparation and use of dyes remained more of an art than a science. The last century has seen not only the discovery of a vast range of synthetic dyestuffs but a great increase in our understanding of the dyeing process. One recent result of this has been the development of a new and important class of dyestuffs—the reactive dyes, so called because they react chemically with hydroxyl groups of the fibres being treated. The ideal reactive dye has yet to be found, but the immediate success of this new class has greatly stimulated research.

Until little over a century ago, dyeing was carried out on small batches of material, using empirical methods of applying extracts of plants, woods, and insects. The expansion of textile production that was so significant a feature of the industrial revolution resulted in the rationalization of textile dyeing and its establishment as an advanced industrial technique. The discovery of the first synthetic dye in 1856 by W. H. Perkin, and the subsequent discovery by P. Griess of the diazo reaction, so important in the development of synthetic dyes, provided a basis for the very rapid growth of dye manufacture. The years following these early discoveries saw the preparation of a vast number of synthetic dyes, and today over 3500 separate chemicals, of which 2000 contain an azo group, are marketed as dyes. The volume of research involved is indicated by the fact that these represent a selection from about one million speculative preparations made in industrial research laboratories. G. S. J. White [1] estimated that in 1958 world expenditure on dyes was about £300 million per annum, and this tremendous quantity of colour is consumed by dyehouses, which in some cases produce as much as ten million yards of dyed cloth a week.

Certain types of organic chemical structure have been found to be of special significance in dye chemistry. For example, substances containing two or more azo links have been found to be very valuable for cheaply dyeing cellulosic fibres such as cotton and rayon, though with only moderate fastness. Where a particular structure has been found to be of special value, groups of dyes that can be applied in the same way and that exhibit the same general resistance to wet treatments have been developed. It is common practice to refer to direct cotton dyes, vat dyes, azoic dyes, and so on, thereby denoting dyes having similar general properties.

The importance of cellulose fibres as textiles has resulted in the development of some eight classes of dyes for them. The most recent addition to these has been the reactive dyes; these have developed more rapidly in use, and aroused more interest, than any previously discovered. Despite the fact that it is still a relatively small group, numbering just over a hundred named products, it is the subject of more research and patent activity than any other. There are two main reasons for this, which are best explained by a general consideration of the dyeing process itself.

When a textile is immersed in a solution of dye, the solution may be merely imbibed or the dye may transfer from the solution to the fibres. Dyeing properties are conferred by the presence in the dye molecule of certain chemical arrangements or groups that encourage either solution in the fibre or the formation of hydrogen bonds or electrovalent links with it. All classes of dyes are absorbed on all fibres by this basic mechanism. The second important property of a dye is its mobility in the fibre; this determines the resistance of the dyeing to washing or other wet treatments. This resistance may be influenced by the formation of hydrogen bonds, electrovalent links, aggregation of the dye in the fibre, or in other ways. In general, dyes that are held only by weak links, such as hydrogen bonds, do not show a high degree of wet fastness, because of the inherent reversibility of the mode of attachment.

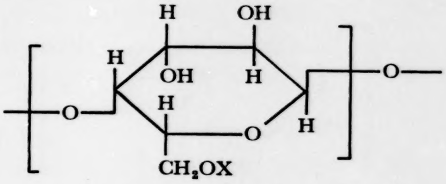
To produce a high degree of wet fastness, it is necessary to bring about some change in the nature of the dye after it has been absorbed by the fibre, so that it differs from its original form. One way is to make the dye insoluble. For example, vat dyes are pigments containing quinone systems, which are first made soluble by alkaline reduction to their sodium salts. After absorption on the fibre, the reduced dye is oxidized to form the insoluble pigment, which is retained mechanically and is highly resistant to washing. Alternatively, the first stage may involve the use of a substance which is not itself a dye but is made to react with another agent, such as a diazonium salt, to form a pigment in the fibre. This last method has been highly developed; it is now even possible to synthesize copper phthalocyanine derivatives on the fibre in this way [2].

Another possible method of permanently retaining the dye in the fibre, but one only recently achieved, is to use a compound that contains labile atoms or groups able to react with the fibre substance itself through a covalent bond; this is the basis of the reactive dyes. Before their advent, it was possible to produce fast dyeings on cellulosic fibres only by means of one of the 'pigmentation' techniques outlined above. To produce a dye that will work in this way, it is necessary to incorporate into the molecule particular groups, such as the quinone group, and this severely limits the dye chemist in developing new shades. In addition, making the dye insoluble results in duller tones. The reactive dyes attach themselves to the fibre by a group that can be separate from the chromophoric system; consequently a very much wider range of shades, including brilliant shades previously unobtainable on cellulosic fibres with any useful degree of fastness, can be obtained.

The production of fast dyeings on cellulosic fibres by a 'pigmentation' technique involves treatment after the initial dyeing process, possibly with several stages. For example, with vat dyes the sequence of operations is reduction→dyeing→oxidation→soaping. The application of reactive dyes is very much simpler, and involves, at most, two stages.

The popularity of reactive dyes thus springs from two characteristics, and their clear technical advantages have caused the first reactive dyes for cellulose—introduced in 1956 by Imperial Chemical Industries under the trade name Procion—to be joined rapidly by other products such as Cibacron (Ciba), Remazol (Hoechst), Drimarene (Sandoz), and Reactone (Geigy). Although the full extent of trade reaction cannot have been foreseen by early workers in the field of coloured cellulose derivatives, some of the potential advantages of this mode of dyeing have long been realized, and the commercial development of reactive dyes for cellulose must be regarded as being, in a sense, the culmination of researches that have been going on for sixty years or more.

Work on the production of coloured cellulose derivatives has been confined to esterification or etherification reactions, generally at the hydroxyl group attached to the 6-carbon in the anhydroglucose unit

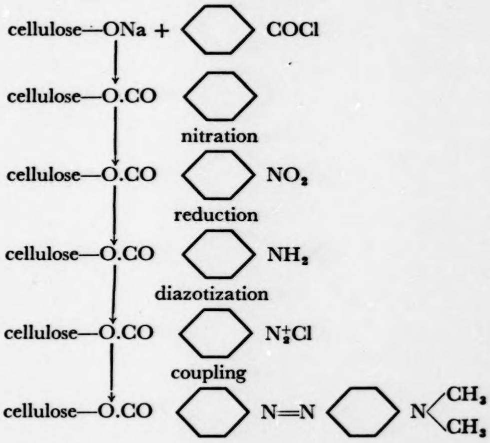


where X is an acyl or aroyl group in the esters and an alkyl or aryl group in the ethers.

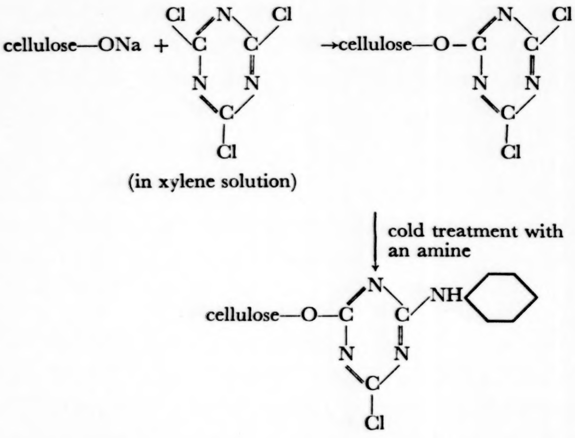
Of the two, the esters of cellulose are the more easily formed; their stability varies with the strength of the esterifying acid, the stronger acids giving the less stable esters. The ethers are in general much more difficult to form, but are more stable. This is not an entirely clear-cut difference, however, for the stability of the product depends on the nature of the ether or ester formed. Because of their relative ease of preparation, attention has naturally tended to be concentrated on the esters.

EARLY RESEARCHES

Among the earliest workers in this field were C. F. Cross and E. J. Bevan [3], who treated cellulose with strong caustic soda solution and then carried out the following sequence of reactions on the 'soda cellulose' so formed:

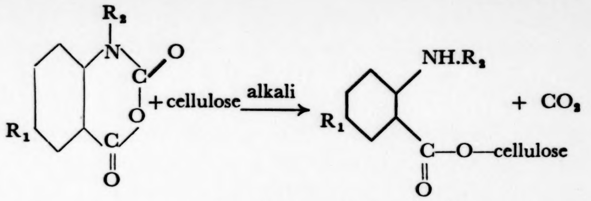


Although the colour of the cellulose derivative so produced was very fast to washing, the nature of the textile was so much altered that it was of little practical value. This approach was very thoroughly explored by many other workers [4–7] using alternative acylating agents, with the aim of shortening the process. From the point of view of subsequent developments, the investigations of Haller and others [8] in 1930–33 are of interest. In this work, the use of 2:4:6-trichloro-1:3:5-triazine for acylating cellulose was examined, using the well-known stepwise reaction-behaviour of this compound; this enables the three chlorine atoms to be reacted separately in the following sequence:



Coloured cellulose may be obtained from the last of these derivatives by coupling with a diazonium salt. However, the reaction procedure used did not depart fundamentally from that of the earlier work. It involved the use of organic solvents, and gave a considerably degraded cellulose of little value for textile purposes.

The dominant feature of all the early work was the use of very strong caustic soda solutions to prepare soda cellulose; it was assumed that cellulose was relatively inert and required vigorous conditions to produce an adequate reaction. This seemed a logical consequence of the work on the preparation of cellulose derivatives for textile and other purposes, where it is necessary to achieve a reasonably high degree of substitution in the cellulose chain. From the point of view of dyeing, however, the degree of substitution required is rarely as much as 1–2 per cent of the total number of hydroxyl groups present, and once this was realized a practicable process became reasonably possible. The possibility was confirmed by F. Günther [9], who was able to esterify cellulose under relatively mild conditions, as shown at the top of the next column.

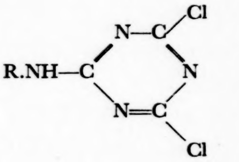


DEVELOPMENT OF PRACTICAL ESTERIFYING DYES

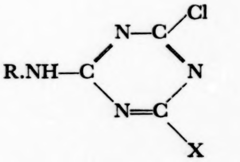
However, until the 1950s there was no extension of Günther's work to cover the use of more satisfactory esterifying agents. I. D. Rattee and W. E. Stephen then examined the use of 2:4-dichloro-1:3:5-triazin-6-yl-amino dyestuffs for the production of coloured cellulose derivatives under the mildly alkaline conditions that are suitable for a dyeing process. This work showed that, if the dyes were suitably forced on to the fibre from the aqueous solution, they were then fixed rapidly. Although the degree of substitution was not high, the efficiency of the reaction was sufficient to provide the basis for attractive dyeing processes yielding fast dyeings over a satisfactory range of shades and depths of shade [10].

It was shown that reactions between these dyes and cellulose could be complete in as little as five seconds, using alkali no stronger than sodium carbonate. Furthermore, it was found that, by varying the *p*H or temperature, the reaction rate could be controlled so that the dyes could be applied under all the conditions encountered in normal dyeing practice.

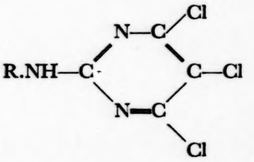
Following these observations, the first commercial range of reactive dyes for cellulose was developed under the trade name Procion. Their success led to intensive industrial research on dyes bearing potential esterifying groups, and this has led to the development of the three following types of reactive esterifying dyes for cellulose:

(1)


Procion dyes (I.C.I.)
In these dyes both chlorine atoms are labile.

(2)


Procion H dyes (I.C.I.)
Cibacron dyes (Ciba)
In these dyes the group X is non-labile.

(3)


Reactone dyes (Geigy)
Drimarene dyes (Sandoz)
In these dyes only one chlorine atom is labile.

COLOURED CELLULOSE ETHERS

Studies of the formation of cellulose ethers resulted in the development of another type of reactive dye which reacts with cellulose fibres under the same general conditions as the esterifying dyes do. These Remazol dyes (Hoechst) [11] are of the general form $R.SO_2.CH_2.CH_2.OSO_3Na$. They form cellulose ethers by elimination of sulphuric acid through the formation of the extremely active etherifying agent $R.SO_2.CH{=}CH_2$.

Coloured ethers have received rather less attention than have the corresponding esters, no doubt because of the relative difficulty of forming them before the discovery of the use of the vinyl sulphonyl group for this purpose. In 1926, D. H. Peacock [12] prepared a fast-dyed cellulose by boiling cotton in an aqueous solution of m-nitrobenzyl-dimethylphenyl ammonium chloride, followed by reduction, diazotization, and coupling. Fibres that could be diazotized were also prepared by C. Granacher [13] by superficial etherification of anhydrous alkali cellulose, using p-nitrobenzyl chloride and similar chloromethyl compounds, or using dinitrohalogenobenzenes with subsequent reduction. Similar experiments were carried out by other workers [6, 14, 15, 16]. The first direct use of dyes themselves for etherifying was made by J. D. Guthrie [17], who found that fast dyeings could be produced by impregnating cotton with sulphate esters of alcohols in caustic soda solution and baking at 100–110° C for an hour or more.

As with the esters, none of this early work led to a practical dyeing process. Only with development of the sulphato ethyl sulphonyl dyes in 1944 [11, 18] and their subsequent application to cellulose, using the reaction conditions previously established in the cellulose ester field [10], was success achieved.

REACTION KINETICS

The four reactive-dye systems currently available are similar in that they all involve reaction with cellulose hydroxyl groups and with water under alkaline conditions, and that in each case a single reaction is decisive in determining the reaction kinetics. This is true of the behaviour of the potentially bifunctional 2:4-dichloro-1:3:5-triazin-6-yl-amino dyes, as well as of the other dyes that are clearly mono-functional, owing to the stepwise character of the reactions of dihalogenotriazinyl compounds with amino or hydroxy compounds. The reaction with water of one of the chlorine atoms in these dyes gives rise to the relatively unreactive monochloromonohydroxy compound, so that as far as the kinetics of dyeing is concerned, the reactions of only one chlorine atom need be considered in the first instance. For this reason, the study of the kinetics of the reactions between dichlorotriazinyl dyes and water or cellulose, to which attention is now turned, provides information that is valid for all reactive systems at present available. Indeed, it can be argued that the theory will be valid for any reactive dye stuff applied to any fibre, provided that only reactions with fibre and with water are involved. The theory is therefore of very wide importance.

In examining this system, it is first necessary to consider the side-reaction with water, for the extent to which this occurs during dyeing obviously governs the efficiency of the process. Further, as will be demonstrated later, the rate of reaction with water at a given *p*H indirectly measures the reactivity of the dye with cellulose.

The rate of hydrolysis of a dichloro-s-triazinyl dye in water can be measured by a number of methods, but since the hydrolysis of the first chlorine atom only is under consideration, a colour reaction between dichloro-s-triazinyl dyes and pyridine in the presence of caustic soda provides an elegant method of analysis [19, 20]. Using this technique, the reaction constants for a number of dichloro-s-triazinyl dyes with water have been determined [21]. A wide variation in reaction rate has been demonstrated, and it has been shown that the reaction rate depends on *p*H.

At *p*H 10, the reactivity with water of the most reactive dye among those commercially available is eighteen times that of the least reactive. To determine whether this is true also of cellulose presents difficulties, and for this reason initial experiments on the reaction of the dyes with hydroxylated substances were carried out with model compounds, such as sorbitol,

$$CH_2OH(CHOH)_4CH_2OH.$$

The sorbitol–dye reaction product itself gives a colour reaction with pyridine in caustic soda, so that the course of the reaction must be followed electrometrically. By maintaining the *p*H at 10·5 or less, at room temperature, and using excess sorbitol, pseudo-first-order reaction constants have been determined; the values are shown in table 1. The wide spread of reactivity with both water and sorbitol can be seen, but the relative constancy of the ratio of reaction constants is also evident.

TABLE I

Reaction constants at pH 10 at 25° C

Dichloro-s-triazinyl dye	k_{H_2O} min^{-1}	$k_{sorbitol}$ min^{-1}	$k_{sorbitol}/k_{H_2O}$
Procion Yellow R	$6{\cdot}0 \times 10^{-4}$	$2{\cdot}8 \times 10^{-2}$	47
Procion Brilliant Yellow 6G ..	$3{\cdot}4 \times 10^{-3}$	$1{\cdot}4 \times 10^{-2}$	41
Procion Brilliant Red 2B.. ..	$1{\cdot}1 \times 10^{-2}$	$4{\cdot}6 \times 10^{-1}$	42
Procion Brilliant Red 5B.. ..	$4{\cdot}6 \times 10^{-3}$	$1{\cdot}8 \times 10^{-1}$	39
Procion Blue 3G	$3{\cdot}7 \times 10^{-3}$	$1{\cdot}4 \times 10^{-1}$	38
Procion Brilliant Blue R	$1{\cdot}1 \times 10^{-3}$	$5{\cdot}1 \times 10^{-2}$	46
Procion Scarlet G	$6{\cdot}1 \times 10^{-4}$	$2{\cdot}9 \times 10^{-2}$	44
Procion Brilliant Orange G ..	$8{\cdot}4 \times 10^{-4}$	$3{\cdot}6 \times 10^{-2}$	43

The mannitol/water system has also been examined over a wide range of alkalinity, using a single dye. The pseudo-first-order reaction constants for mannitol and for water have been measured, and it has been shown that the ratio $k_{mannitol}/k_{water}$ is reduced as the alkalinity is increased. However, this is not necessarily a complete picture of the true situation, since mannitol ions can be formed under the *p*H conditions of the experiment. Reaction with these ions is possible, and that it does occur is strongly indicated by the fact that when the bimolecular reaction constants are calculated on the basis of mannitol O^- and OH^- concentrations, using known *p*K values, the ratio of the two reaction constants proves to be constant at all *p*H values, as shown for a typical dye in table II.

TABLE II

Comparison of pseudo-first-order and bimolecular reaction-constant ratios (mannitol/water system)

[NaOH]	$k_{mannitol}/k_{water}$	$k_{mannitol\ O^-}/k_{OH^-}$
0·1N	1·05	13·5
0·2	0·89	13·5
0·4	0·66	14·3
0·6	0·58	14·7
0·8	0·40	12·2
1·0	0·41	14·5
2·0	0·20	15·8

Mean value $k_{mannitol\ O^-}/k_{OH^-} = 14{\cdot}1 \pm 1{\cdot}2$

This result shows that the reaction is governed by the concentration of carbohydrate ion present, and that for a given carbohydrate, the various dyes all show, despite their wide variation in reactivity, the same ratio of reactivity with the carbohydrate ion as compared with the hydroxyl ion.

Attention may now be turned to the reaction with cellulose itself. The heterogeneous character of the cellulose/water/dye system complicates the application of the above reaction mechanism to the dyeing of cellulose, because calculation of concentrations of the various reactants must involve consideration of affinity and diffusion, and because there may be steric factors in the solid phase to be considered. In addition, it is necessary to demonstrate that cellulose ions can exist under given reaction conditions. Fortunately, this kind of problem has arisen in other fields of work, and P. V. Danckwerts [25] has evolved an equation that may be simplified to a form suitable for dyeing phenomena. His equation is

$$Q=c\left(t+\frac{1}{2k}\right)\sqrt{(dk)}$$

where Q = the amount of dye diffusing through unit area of solid surface
c = the equilibrium surface concentration
d = the diffusion coefficient
k = the reaction constant for cellulose
t = the duration of dyeing

This equation has been applied by H. Sumner [22] to the behaviour of reactive dyes on cellulose. As the Danckwerts equation refers to the ideal case of diffusion into a slab of infinite thickness, it was necessary to carry out experiments with viscose film rather than fibres. By determination of the values of Q, c, and d, Sumner was able to calculate values of k over a range of *p*H and to compare these with the reaction rates with water. As in the mannitol/water system, the ratios are not constant and vary between 25 and 10. However, the behaviour of the mannitol/water system suggests that it is in fact the values of k based on reaction with ionized hydroxyl groups that ought to be compared.

Sumner [23] has given strong evidence, based on a Donnan equilibrium and S. M. Neale's value for the *p*K of cellulose [24], supporting the theory that sodium hydroxide is absorbed by a neutralization mechanism. Excellent agreement has been shown between observed values of caustic soda uptake and values predicted on the basis of the neutralization theory. When the reaction constants determined by the use of the simplified

Danckwerts equation are adjusted to allow for [cellulose O^-], and the ratios of the new reaction constants with those for [OH^-] are calculated, the latter are found to be almost constant at all alkaline *p*H values (table III).

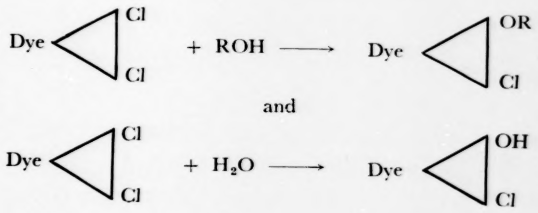
TABLE III

Comparison of pseudo-first-order and bimolecular reaction-constant ratios (cellulose/water system at 20° C)

*p*H	k_{H_2O} min^{-1}	$k_{cellulose}$ min^{-1}	$k_{cellulose}/k_{H_2O}$	$k_{cellulose\,O^-}/k_{[OH]^-}$
8·00	$9\cdot12 \times 10^{-4}$	—	—	—
8·81	$3\cdot63 \times 10^{-3}$	—	—	—
9·14	$5\cdot62 \times 10^{-3}$	—	—	—
9·46	$6\cdot92 \times 10^{-3}$	$1\cdot72 \times 10^{-1}$	25	0·95
9·74	$1\cdot29 \times 10^{-2}$	$3\cdot30 \times 10^{-1}$	26	1·16
10·99	$2\cdot82 \times 10^{-2}$	$7\cdot03 \times 10^{-1}$	25	1·16
11·70	$6\cdot46 \times 10^{-2}$	1·59	25	1·28
11·94	$1\cdot00 \times 10^{-1}$	1·87	19	1·06
12·63	$4\cdot90 \times 10^{-1}$	7·71	16	1·37
12·73	$6\cdot03 \times 10^{-1}$	6·52	11	1·03
		Mean value = 1·14 ± 0·14		

From these experiments an extremely simple picture emerges of the reaction between reactive dyes and cellulose. Reaction takes place with hydroxyl ions and with ionized cellulose, and since the concentration of each of these is extremely small in the absence of alkali, the rate of reaction is negligible. The addition of alkali increases the rate of both reactions, but the preferential reaction with the cellulose cannot be explained on the basis of an inherently greater reactivity of cellulose, because both reaction constants (that is, for cellulose O^- and for OH^-) are of the same order of magnitude. Other factors must therefore be considered. Firstly, there is the affinity of the dye for the cellulose fibre, which causes the dye concentration in the fibre to be as much as 500 times greater than that in the water. Secondly, there is the lower dissociation constant for cellulose as compared with water (*p*K 13·7 as against *p*K 15·7), giving, at any given *p*H, a greater degree of ionization in the cellulose relative to the water. These two factors are sufficient to outweigh the considerable preponderance of water in the practical dyebath. The Danckwerts equation suggests that the affinity of the dye for the fibre, expressing itself through the value of *c*, is the most decisive factor, because the other influences, diffusion and reactivity, are expressed as square roots in the equation. However, high affinity is frequently associated with low diffusion, and some balance between the two is necessary in a useful dye.

The reaction constants determined by Sumner for the reaction of cellulose with the reactive dye are unlikely to refer to a single dye-fibre reaction. T. S. Gardner and C. B. Purves [26] have shown that toluene-p-sulphonyl chloride reacts with the three hydroxyl groups of cellulose at C_6, C_2, and C_3 in the ratio 23·4 : 2·16 : 0·106. There is therefore the strong probability that in Sumner's work the esterifying reaction involved both primary and secondary hydroxyl groups. Since all three possible reactions are of the same order, this will be concealed in $k_{cellulose}$ determinations and will in no way invalidate the results. Because it is not possible to determine the actual sites in the cellulose molecule at which reactive dyes attach themselves, this has been studied indirectly, using dichlorotriazinyl amino dyes and model compounds. The rates of reaction of a number of such dyes with n-propanol and isopropanol have been determined [19]. Because the dyes are insoluble in the pure alcohols, the experiments were carried out in 50 per cent solutions of the alcohol in water, and a correction for the reaction rate in water alone was applied. This device assumed that the reactions



(where ◁ represents the s-triazinyl amino residue and R represents the aliphatic residue) proceed independently.

For Procion Red G the first-order reaction

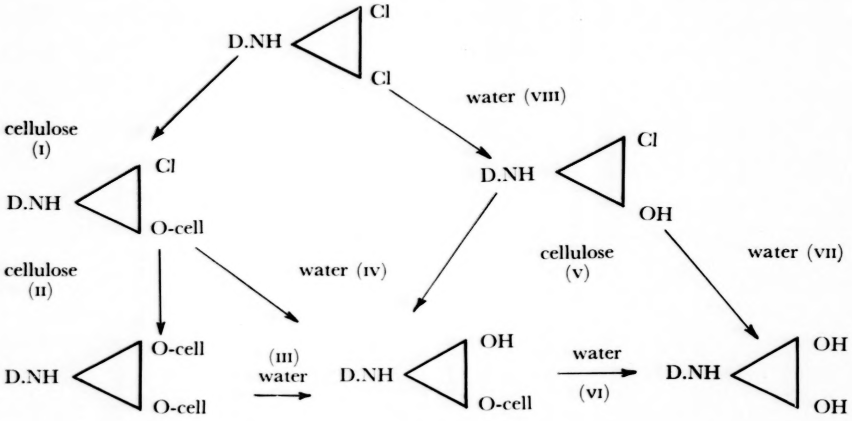
TABLE IV

Dye	Ratio of reaction constants (primary/secondary)
Procion Brilliant Red 2B	3·7
Procion Brilliant Blue R	6·6
Procion Red G	7·2
Procion Yellow R ..	7·2
Procion Blue HB* ..	13·4
Procion Brilliant Red H3B*	15·2

* Monochlorotriazinyl dyes.

constants ($k(\mathrm{min}^{-1})$) were found to be 0·58 for n-propanol and 0·08 for isopropanol. The ratio of the two reaction constants is thus in moderate agreement with the ratios observed by Gardner and Purves. It is to be expected that the ratio will vary with different acylating agents, and this is well demonstrated by the following comparison (table iv) of the reaction-rate ratios for different dichlorotriazinyl dyes.

Reactive dyes bearing other reactive groups behave similarly. These are, however, different from the dichlorotriazinyl dyes in that they are not only less reactive but also possess a single reactive centre. The dichlorotriazinyl dyes, on the other hand, can pursue a multiplicity of reactions, as shown below,



where D is the dyestuff residue and ◁ is the s-triazine residue. Reaction i has been discussed; reaction v is very slow and may be virtually ignored; and reaction vi does not occur under conditions encountered in dyestuff application. Thus the rate-determining reactions are i and viii, and the picture of the reaction kinetics previously drawn remains true. Reactions ii, iii, and iv are also of interest, however, because they give rise to dye-cellulose reaction products having different properties.

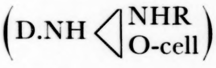
TABLE V

	Acid hydrolysis	Alkaline hydrolysis
Type 1 ..	Relatively stable, slowly converts to type 3.	Initially converts to type 2 and ultimately to type 3.
Type 2 ..	Relatively stable	Relatively stable, ultimately converts to type 3.
Type 3 ..	Relatively unstable	Relatively stable

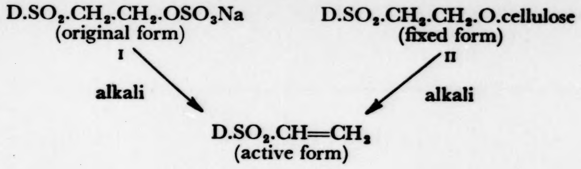
Cross-linking (reaction ii) and subsequent cross-link rupture have been demonstrated by numerous workers [27–29]. In practical dyeing three types of dye-cellulose product may be formed: these may be designated type 1, type 2, and type 3, arising from reactions i, ii, and iii or iv respectively. Because of the stepwise reactivity of the halogens on halogeno-s-triazinyl compounds, it is possible to produce dyeings that are almost perfect examples of the three types, and the properties of these with respect to acid or alkaline hydrolysis have been studied [21]. These are summarized in table v.

LIMITATIONS OF REACTIVE DYES

The above results have practical significance, because the fastness of some dichlorotriazinyl dyes changes during storage of dyed materials. For example, type 3 dyeings of a few dichlorotriazinyl dyes gradually lose their high initial fastness, and fabrics dyed with them 'bleed' when put into water; this is because the dye-fibre bond ruptures. As types 1 and 2 of the same dyes change slowly to type 3 during storage, and more rapidly if treated with alkali, they too deteriorate, though not so quickly. This effect is offset industrially in two ways. Firstly, since not all dyes show the effect, it can be minimized by careful selection during the initial investigation. Secondly, use can be made of the remaining reactive halogen in type 1 dyeings by causing it to react with an amine to produce what is known as a type 4 dyeing


$\left(\mathrm{D.NH}\lhd\begin{matrix}\mathrm{NHR}\\ \mathrm{O\text{-}cell}\end{matrix}\right)$, which is relatively stable.

This problem is encountered to a greater or lesser extent with all acylating dyes, that is, those based on s-triazine or 1 : 3-diazine. The dyes that form vinyl sulphone are free from this trouble, but possess defects of their own. The compound $\mathrm{D.SO_2.CH_2.CH_2.O\text{-}cellulose}$ formed by the reaction of these dyes with cellulose is related to their original form, $\mathrm{D.SO_2.CH_2.CH_2.OSO_3Na}$, if cellulose is regarded as a weak acid. Under alkaline conditions causing ionization of cellulose, the dye-fibre compound tends to revert to the vinyl sulphone, as shown overleaf.



Cellulose, being a weaker acid than sulphuric acid, requires a higher *p*H to bring about this reaction. However, the result is that under hot alkaline washing conditions, hydrolysis occurs, leading to the stripping of dye from the fibre, and, because of liberation of the active form of the dye, there is some reaction with adjacent white cellulosic material, giving rise to a stain.

Clearly the ideal reactive dye for cellulosic materials has yet to be found, and intensive research on a world-wide scale is continuing. The economic basis for these efforts is the quite remarkable success achieved by this new class of dyes due to their very wide gamut of shades, simplicity of application, and not least their suitability for numerous advanced techniques of application, leading to quite startling increases in productivity.

REFERENCES

[1] White, G. S. J. *J. Soc. Dy. Col.*, **76**, 16, 1960.
[2] Gund, F. 'Symposium on Textile Printing', p. 147. Society of Dyers and Colourists. 1953.
[3] Cross, C. F. and Bevan, E. J. 'Researches on Cellulose', p. 34. 1895.
[4] Briggs, G. J. *Z. angew. Chem.*, **26**, 256, 1913.
[5] Schroeter, G. *Ber. dtsch. chem. Ges.*, **39**, 1570, 1906.
[6] Dirklage, R. British Patent Specification 389279.
[7] Riesz, E. *Bull. Soc. industr. Mulhouse*, **99**, 349, 1933.
[8] Ciba. British Patent Specification 363897.
[9] Günther, F. U.S. Patent Specification 1567731.
[10] Alsberg, F., Rattee, I. D., Stephen, W. E., Marshall, W. J., Speke, R. W., and Weston, C. D. British Patent Specification 797946.
Alsberg, F., Heslop, R. N., Rattee, I. D., Stephen, W. E., Marshall, W. J., Speke, R. W., and Weston, C. D. British Patent Specification 798121.
[11] Farbwerke Hoechst. British Patent Specification 733471.
[12] Peacock, D. H. *J. Soc. Dy. Col.*, **42**, 53, 1926.
[13] Granacher, C. British Patent Specifications 346385, 347117, 347263.
[14] Dreyfus, H. British Patent Specification 344420.
[15] I.G. Farben. German Patent Specification 492062.
[16] Pancirolli, F. *Boll. Rep. Fibre tess. Veg.*, **32**, 314, 1937.
[17] Guthrie, J. D. *Amer. Dyest. Rep.*, **41**, 13, 30, 1952.
[18] Ciba. British Patent Specification 587467.
[19] Dawson, T. L., Preston, C., and Fern, A. S. *J. Soc. Dy. Col.*, **76**, 210, 1960.
[20] Sause, R. (Unpublished work.)
[21] Preston, C. and Fern, A. S. *Chimia*, **15**, 177, 1961.
[22] Sumner, H. *J. Soc. Dy. Col.* (In preparation.)
[23] *Idem. Ibid.*, **76**, 672, 1960.
[24] Neale, S. M. *J. Text. Inst.* **20**, T373, 1929.
[25] Danckwerts, P. V. *Trans. Faraday Soc.*, **46**, 300, 1950.
[26] Gardner, T. S. and Purves, C. B. *J. Amer. chem. Soc.*, **64**, 1539, 1942.
[27] Vickerstaff, T. *Amer. Dyest. Rep.*, **47**, 33, 1958.
[28] Wegmann, J. *Melliand Textilber.*, **39**, 1006, 1958.
[29] Schwertassek, K. *Faserforsch. u. Textiltech.*, **9**, 321, 1958.

Greek fire

J. R. PARTINGTON

The fall of Constantinople in 1453 marked the end of the Middle Ages and was an event of profound historical significance. Before this, Greek fire had proved of great value to the Byzantine empire in its centuries-long resistance to a host of enemies. Despite the military importance of this incendiary substance, which seems to have been due as much to its terrifying appearance as to its destructive powers, its chemical nature has long been controversial. In this article convincing evidence is advanced that Greek fire was neither gunpowder nor a mixture of quick-lime and sulphur, but a light petroleum fraction thickened by addition of resin or sulphur. The traditional attribution of the discovery to Kallinikos is not conclusive.

The use of some form of fire in combat goes back to very early times. The Assyrians used burning pitch or crude petroleum in defending the walls of cities in siege warfare. Fireships were used by the Greeks from the fifth century B.C., but incendiary arrows, still used by undeveloped races, came into common use in Greek armies only after the death of Alexander the Great. Appollodorus of Damascus (A.D. 60–130), Hadrian's architect, describes an elaborate apparatus that used bellows for throwing burning charcoal powder on to stone walls, which were then cracked by splashing them with vinegar. Roman troops used incendiary arrows tipped with a mixture of sulphur, resin, bitumen, and petroleum; large spears with iron heads tipped with this incendiary mixture were thrown in sieges by engines mounted on wooden towers.

Petroleum was certainly well known in antiquity, and was obtained from Parthia, Media, Persia, and the region around the Caspian and Black Seas. Many of the sources now yielding petroleum in these regions were known and worked in quite early times; some which are mentioned in ancient accounts have been rediscovered fairly recently. The name *naphtha* used by the Greeks, is derived from the old Persian name *naft.* Strabo reports that when Alexander the Great was in the city of Ecbatana in Media he was shown a demonstration in which naphtha was sprinkled on a street and ignited at one end, when the flame flashed in an instant to the other end. Pliny tells us that liquid bitumen from Babylon was burnt in lamps. There was a 'white naphtha' which was perhaps the product of filtering cruder material through adsorbent earth or, soon after the beginning of the Christian era, of distillation. Oil of turpentine and mercury were prepared by a crude form of distillation in which the vapour from a material heated in a pot was condensed on a cooled lid; but true distillation, in which the still is separated from the receiver by a tube, was an invention attributed to Mary the Jewess, an Egyptian alchemist of the early Christian era. From Egypt, distillation apparatus passed to Syria and the Arabs, and to Constantinople, where chemistry was taught by Stephanos in the first half of the seventh century. Glass still-heads of the fifth to eighth centuries, excavated in Syria, are of the same form as those shown in Greek alchemical manuscripts and attributed to Mary. Crude petroleum, probably from the Caspian region, was imported into Constantinople and used on a very large scale for heating baths. Although we have no definite information about the distillation of petroleum it seems very likely, in view of the history of the process and the availability of the raw material, that light petroleum was obtained by distillation before the eighth century.

A different kind of incendiary material was that called 'automatic fire'. It was composed of quicklime mixed with easily inflammable materials such as sulphur, resins, powdered pyrites, and petroleum. In contact with water, the heat evolved by the hydration of the quicklime, it was supposed, would be sufficient to ignite the mixture, which was to be put on or under the wooden roofs of buildings, or on wooden catapults, there to be ignited by dew or rain. A recipe for this automatic fire is given in a sixth-century interpolation of a work, *Kestoi*, written in Greek by Julius Africanus, who was born about A.D. 170 in Jerusalem and was an engineer in the service of the Emperor Alexander Severus. The recipe says the mixture was stored in bronze boxes with airtight lids, and for use was secretly smeared over the engines of the enemy in the evening: 'when the sun rises all will be burnt'. The idea of this automatic fire seems to have come from a story in Livy, saying that in the Bacchanalia of 186 B.C. torches with heads composed of quicklime and sulphur were dipped into water and withdrawn in flames. A Chinese account of a river battle of A.D. 1161 says that 'thunderbolt missiles' consisting of paper bags filled with quicklime and sulphur were thrown on the water, when they burst into flames and leaped upwards. Nevertheless, it seems improbable that such compositions would act effectively or were ever actually used in warfare. The mixture of sulphur and quicklime has been identified with Greek fire, but the properties and mode of use of the latter were quite different.

THE BYZANTINE AND ARABIC CONTRIBUTIONS

True Greek fire was invented in Constantinople, for long centuries a bulwark against Asiatic conquest; the Byzantines called it 'liquid fire', the misleading name *feu grégeois* being coined by Crusaders. The Byzantines did not call themselves Greeks but regarded themselves as Romans, the name Greek at that time being rather a term of reproach. Both Byzantine and Arab chroniclers agree that from the end of the seventh century an incendiary far more terrible than anything known before came into use in sea battles. Byzantine Constantinople was long able by diplomacy and arms to resist attacks by Arabs, Persians, Goths, Huns, Normans, Russians, and Franks; but it fell at last, in 1453, to the Turks, whom it had resisted for centuries.

The Byzantine culture was refined and active—it preserved the heritage of the older civilizations of the West—and the army was well disciplined. The Roman arsenals even in their best days could have supplied no scientific or mechanical contrivances unknown to the Byzantine engineers, whose education, discipline, and practice had been perpetuated in unbroken succession from the times of Trajan and Constantine. The city was full of inventors and craftsmen, and, as we have noted earlier, the alchemist Stephanos lived there a little before Greek fire was invented. There was keen interest in chemistry, and in the seventh and eighth centuries Greek chemical treatises were collected and edited. In the ninth century the philosopher Leo of Thessalonika made for the Emperor Theophilus a golden tree, whose branches carried artificial birds that flapped their wings and sang; a model lion that moved and roared; and a bejewelled clockwork lady who walked. While of no importance in themselves, such contrivances are evidence of the evolution of a considerable degree of technical skill. In such circumstances the invention of Greek fire and of effective means of using it in warfare need occasion no surprise.

Theophanes in his *Chronography*, written in 811–15, says that in 671 the Emperor learned that the Arabs were preparing to attack the city and had established their winter quarters in Asia Minor. In reply to the threat he ordered the assembly of a fleet of fireships equipped with 'siphons', fire-projecting devices to whose precise nature we will turn later. In the spring the siege commenced and was carried out during the summer months for seven years. Finally, the Arab boats and sailors were destroyed at Kyzikos, partly by a storm and partly by a form of fire alleged to have been invented by the architect Kallinikos of Heliopolis in Syria, who had fled to the Romans.

The accounts of the sieges given by Byzantine and Arab historians are, as Gibbon and others have pointed out, confused. It seems, however, that there were five Arab expeditions in all—three in 665, 668, and 674; one in 715, when Greek fire was again used; and one in 782. The Arabs only once made a landing, but failed to capture the city; in 782 they again ventured within sight of Constantinople, but again turned away.

The contribution of Kallinikos to these events is uncertain, and the association of his name with the invention of Greek fire may be unjustified. Theophanes says that in 671 fireships were fitted with siphons, and it may be true that Kallinikos provided a new kind of incendiary mixture that made these more effective. Nevertheless, it is clear that at that time Byzantium already contained knowledgeable Greek chemists perfectly capable of making the invention for which Kallinikos received credit in official circles. Nearly two centuries later, the Emperor Constantine Porphyrogenitus (905–59) said that the recipe for the fire had been revealed by an angel to Constantine the Great, and that those making it known were anathema and would die instantly in disclosing it. Porphyrogenitus, who lent troops and military engines to his allies, reserved the secret of its manufacture and sent the fire ready-made to his faithful but not wholly reliable dependants. But Greek fire was not a prerogative of either angels or emperors; it had been discovered by chemists, and further east, in Saracen lands, other chemists were soon busy with its independent discovery.

The way in which Greek fire was used in naval battles is described in a treatise on tactics and strategy attributed to the Emperor Leo, either the Isaurian (A.D. 717–41) or the Armenian (A.D. 886–911). He says: 'Many valuable military engines of destruction have been invented by the ancients and moderns for attacking ships and sailors. Of this kind is the emission of fire with thunder and burning smoke, which burns ships'. The context indicates that 'thunder' does not denote an explosion; the ships were burnt and not blown up. The story has grown in the telling, and later exaggerated descriptions, which reach a climax in Oriental accounts, have misled many into thinking that gunpowder must have been meant.

Leo goes on to speak of 'the other device of the small siphons discharged by hand from behind iron shields, which are called hand-siphons and have recently been manufactured in our dominions . . . these can throw the prepared fire into the faces of the enemy'. The allusion was thought to have been to grenades filled with gunpowder thrown by troops, but their nature is not now in doubt. The term siphon was commonly used for a pump, and a small hand-pump projecting Greek fire is shown in a tenth-century Byzantine manuscript. In the next century Abu'-l-Qāsim, after describing and depicting in his book on surgery a cylindrical syringe with a piston, says it is 'like the tube by means of which naphtha is thrown in sea-combats'.

Leo further describes how the Greek fire was projected from a ship: 'The front part of the ship had a bronze tube so arranged that the prepared fire could be projected forward, to left or right, and also made to fall from above. This tube was mounted on a false floor above the deck on which the specialist troops were accommodated, and so raised above the attacking forces assembled in the prow. The fire was thrown either on the enemy's ships or in the faces of the attacking troops'. In this case also we have a picture, although not a very accurate one, of the contrivance in a tenth-century manuscript (figure 1). There seems to be no doubt that the essence of it was that a flaming liquid was projected by a force-pump through a metal tube.

Anna Komnena, daughter of the Byzantine Emperor Alexios I Komnenos, describes a naval battle between the Greeks and Pisans near the island of Rhodes in 1103. Each of the Byzantine galleys was fitted in the prow with a tube ending with the head of a lion or other beast made of gilded brass or iron, frightful to behold, through the open mouth of which fire could be projected by means of a flexible apparatus. The technique was to ram an enemy ship in the stern and to pump the fire over it. The Pisans fled, having no previous experience of this device, wondering that fire, which usually burns upwards, could be so directed downwards or towards either side, according to the will of the engineer who discharged it. This account agrees exactly with that given earlier by the Emperor Leo. Apparently the projection apparatus was made to swivel, the open mouths of the fixed beasts being wide enough to allow for this.

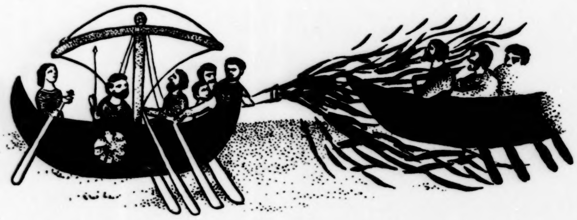


FIGURE 1 - *Use of Greek fire in a naval battle.* (*From a tenth-century Byzantine manuscript.*)

Anna also mentions an incendiary blowpipe, in this case apparently fed with a powder or paste projected by the breath past a kindling device on the tip of the tube; it seems to have done little damage, though the effect was formidable. She says that in an encounter in an underground sap, at the siege of Durazzo (1108), the Normans had their beards singed by the Byzantines but were not otherwise much harmed.

After their experience before Constantinople, the Arabs, to whom petroleum was quite familiar, soon made Greek fire for themselves with the help of their own very expert chemists. Under the Abbasid rulers, of Persian origin, the Muslim armies were equipped with 'naphtha troops' who wore fireproof suits and projected the burning liquid by special pumps. In the tenth century a 'fierce fiery oil' was brought to China by Arabs, and soon after a special apparatus was used by the Chinese to project it by a pump.

Richard I, in his voyage from Cyprus to Acre, captured a Saracen transport ship laden with all kinds of weapons and an abundance of Greek fire in bottles, which an eye-witness had seen put aboard at Beyrut. Pictures of such ships, stacked with spherical pots or bottles, are to be seen in a thirteenth-century Arabic manuscript. In the twelfth century, surprisingly large quantities of petroleum were available to the Arabs: 20 000 barrels of it were used to burn down the city of Cairo to prevent its recapture by the

Franks. Arabic accounts refer to the large stocks of naphtha held also by the Byzantines, who were alleged to store it in churches.

To the Crusaders, the Greek fire used by their Saracen opponents was something quite new and unexpected. The siege of Mansura in the Nile Delta in 1249, in which King Louis IX of France was captured by the Mamluk Sultan, is described by Jean de Joinville (1224–1319), in whose time gunpowder was known and cannon were imminent. The Saracens discharged the fire, burning in tubs, from catapults: 'It was like a big tun and had a tail of the length of a large spear: the noise which it made resembled thunder and it appeared like a great fiery dragon flying through the air, giving such a light that we could see in our camp as clearly as in broad day'. When the French commander saw the Saracens preparing to discharge the fire he announced in panic that they were irretrievably lost.

The Crusaders believed that anyone struck by Greek fire was doomed, not only because it could kill instantly, but because its 'fall-out' was a deadly poison, as even trifling gunshot wounds were later supposed to be. Yet among all the terror Joinville does not produce a single fatal casualty due to Greek fire. Clearly its effects were greatly exaggerated. There are some interesting Arabic treatises on incendiary warfare of the twelfth and thirteenth centuries which contain highly coloured accounts of the use of distilled naphtha. It was said that if it was thrown in jars against a castle, the exhalation of the mixture would inflame and the castle would crumble with a noise of thunder and a frightful whistling. Any debris left could be destroyed by a prepared water and naphtha, which burned with a thick black smoke, lethal by its terrible and suffocating stench. That such gross exaggeration could be current is evidence of the terror that Greek fire aroused and of the moral advantage to its possessors.

THE COMPOSITION OF GREEK FIRE

The nature of Greek fire, which has been the subject of research and argument for a long time, seems to follow almost automatically from the above descriptions of its use and effects. It was a liquid, prepared by some process, that was either projected burning by pumps, especially in sea warfare, or in land warfare was commonly thrown in hand grenades or, burning in large tubs, by catapults. The pumps throwing it were called siphons; hand siphons were small pumps carried by soldiers.

Jacques de Vitry and Gossuin of Metz, in the period of the Crusades, significantly say that Greek fire was made from a liquid obtained from springs in the East, and this supports the view to which we have already briefly alluded, that its basis was a light petroleum fraction obtained by distillation. Petrol by itself would not be very effective when used in flame-projectors, as the jet would dissipate too rapidly and would not carry far enough. To overcome this technical defect it was probably thickened almost to a jelly by dissolving in it resins and perhaps also sulphur. Such a glutinous mixture would cling to its target and so be more destructive. The particular mixture, or mixtures, used and the mechanical means of projecting it, together constitute the invention of Greek fire.

The composition of Greek fire may be further inferred from some accounts compiled long after its first use. One recipe is given in the so-called 'Book of Fires of Mark the Greek', usually called the *Liber Ignium*. Although this has long been supposed to be a Latin translation of a Greek original, as the name suggests, I have shown that it is practically certain that no Greek original existed, and that the document was compiled in Spain, between 1250 and 1300, probably by a Jew, from Arabic sources. The recipe reads: 'You will make Greek Fire in this way. Take live sulphur, tartar, sarcocolla and pitch, boiled salt [*sal coctum*], petroleum oil and common oil. Boil all these well together. Then immerse in it tow and set it on fire. Then kindle the fire, which is not extinguished except by urine, vinegar, or sand'. Although it has been argued that 'boiled salt' might be saltpetre, there is little doubt, as other sources show, that it was common salt refined by boiling with water and evaporation, and the name is in fact never used for saltpetre. The salt would give a yellow colour to the flame and make it look very hot.

Another recipe occurs in an Arabic manuscript composed for Saladin about 1193; it specifies a 'naphtha' for throwing that was composed of petroleum, tar, sulphur, and resin. A third recipe is given in an Arabic manuscript copied from an original of about 1225. According to this, petroleum is first purified and made white by treatment with alum, white of egg, and other substances, followed by filtration through sand. It is distilled in a cucurbit and alembic, and it is noteworthy that, in the process of distillation said to be used in Basra, the distilled naphtha is redistilled in a vessel by burning under it reeds 'one by one', that is, by a very gentle heat, which indicates that a low-boiling petroleum fraction was involved. A mixture of this with resin, fat, black pitch, and sulphur will burn on water.

Anna Komnena gave the following recipe. 'This fire is made by the following arts. From the pine and certain such evergreen trees inflammable resin is collected. This is rubbed with sulphur and put into tubes of reeds and is blown by men using it with a violent and continuous breath.' There is no mention of distilled petroleum, the essential ingredient, but as the daughter of the Byzantine Emperor she would have had good reasons for not revealing the whole secret.

Despite such evidence pointing to naphtha, many have identified Greek fire with gunpowder, but this does not agree with its properties, nor with the way in which it was used. Further, there is no information of any sort which could suggest that saltpetre, the essential constituent of gunpowder, was known at the time when Greek fire was being quite extensively used. Others have suggested different materials, such as the early incendiary materials and the 'automatic fire' composed of quicklime and sulphur, and have assumed that Greek fire contained all these. Being solid, such a mixture could not have been projected as Greek fire was, nor would it have had its properties. Compositions fantastic for the times of which we are speaking, such as calcium phosphide, have also been proposed. All the evidence, however, which I have set forth at length in my 'A History of Greek Fire and Gunpowder' (Cambridge, 1960), seems to indicate that saltpetre was not an ingredient, that Greek fire was quite distinct from gunpowder, and that the basic constituent was naphtha thickened by the addition of resin or sulphur.

Book reviews

SCIENCE AND TECHNOLOGY

McGraw-Hill Encyclopaedia of Science and Technology. *Vols. I–XIV, Aba–Zyt,* 9040 *pp.*; *Vol. XV, Index,* 536 *pp. McGraw-Hill Book Co. Inc., New York; McGraw-Hill Publishing Co. Ltd, London.* 1960. £75 11*s.* 6*d. net.*

There has long been a need for a reference book on science that would be more detailed than the large general encyclopaedias, but would yield its information more easily than a textbook. This need has been felt not only by scientists who wish to follow progress in fields other than their own, but by the general public in a society that attaches increasing importance to science. This encyclopaedia is intended to satisfy this need.

The editors had at the outset to decide on the range of subjects to be included. Their decision was that the basic matter of the natural sciences, and their major applications in engineering, agriculture, forestry, industrial biology, and food and other technologies, should be included. On the other hand, only the 'pre-professional science' of medicine, pharmacy, pharmacology, and the behavioural sciences was to be considered. Another restriction was that history and philosophy were to be included only when they were needed for the basic understanding of a scientific concept or of its technical application.

If its aims are to be fulfilled, there must be some restriction on the content of a book of this kind, and that on the range of subjects is a reasonable one. So, too, is that excluding all but the essential history and philosophy, although here the results are sometimes surprising: the entry under 'atomic bomb', for instance, gives no more than a brief description and explanation of the principles involved, with no hint of the background.

Projects of this kind create great editorial problems which are not simplified if those participating are separated by barriers of distance or language. It is, therefore, not surprising to see that all the sixty-three consulting editors are American. Less satisfactory is the fact that all but sixty of the contributors, rather more than two thousand in all, are also American. This is open to criticism on practical rather than nationalist grounds. So far as the scientific entries are concerned the point is of no great significance, for at the level of these volumes an American contributor would write much the same as any other of comparable calibre. But it does weaken the technological sections, where there tends to be too heavy an emphasis on American practice. For example, the important article on electric power generation makes no mention of practice outside the United States.

However, the choice of authors has, presumably, been governed largely by practical considerations and the need for speed. In this last respect the encyclopaedia is impressively, indeed astonishingly, up to date. The article on the Moon, for example, gives a full account of the photography of the far side, together with an illustration. The discovery of ^{129}Xe in meteorites, made only last year, is given in the article on Meteorites. The work is to be kept up to date by a series of year books, and provision has been made for constant revision of the main text.

The articles on specialized topics are extremely thorough: the texts are lucid and detailed, and the illustrations numerous and clear. Sampling shows that the articles contain, in general, the information expected, although sometimes under somewhat unexpected headings. Here the comprehensive index is very valuable, although one could have wished for an additional classified index. The articles themselves contain cross-references and short bibliographies.

The above criticisms are, however, of a relatively minor nature, such as no work of this magnitude could hope to escape. Generally speaking, the work successfully achieves its aims, and is to be welcomed. TREVOR I. WILLIAMS

THEORY OF MATRICES

Introduction to Matrices and Linear Transformations, *by Daniel T. Finkbeiner II. Pp. vii + 248. W. H. Freeman & Co. Ltd, London.* 1960. 38*s. net.*

There are various methods for presenting the theory of matrices to students, and the course adopted in this book is the modern one of using the techniques of abstract algebra. The first three chapters deal with abstract systems, vector spaces, and linear transformations. The theory of matrices is then developed up to the canonical form for a linear transformation. The last two chapters deal with metric concepts (bilinear and quadratic functions, length, distance, and orthogonality) and with functions of matrices, including an application to differential equations. There are well-written short appendices on algebraic concepts and combinatorial equivalence, and the answers to a number of exercises are given.

Matrix calculus looms large in various branches of applied science, including economics, psychology, and operations research. Students of these subjects who have taken mathematics to the first-year level at the university will find the present book quite readable; for, as the author remarks, 'the major prerequisite . . . is the ability to reason abstractly, to proceed logically from hypothesis to conclusion'.

L. S. GODDARD

ASTRODYNAMICS

An Introduction to Astrodynamics, *by R. M. L. Baker, Jnr., and M. W. Makemson. Pp. xiv + 358. Academic Press Inc., New York; Academic Books Ltd, London.* 1960. $7.50.

This book is really an introduction to the dynamics of the solar system. As such it is welcome, particularly as it is well suited to the needs of those who wish to obtain a detailed understanding of the motion of planetary satellites and space probes but who are not primed with astronomical knowledge. The treatment is by turns mathematical and descriptive. The descriptive sections are quite readable; the mathematical sections assume the common mathematical equipment of the science graduate, without any previous knowledge of dynamics.

The first five chapters are introductory in character. They contain a discussion of the elements of orbital theory, an historical description of the astronomy of the solar system, and conclude with an excellent section on the deduction of the principal units of distance in the solar system, and of the masses of the Moon and planets.

The remaining chapters deal with orbit determination from observations, including the improvement of orbital elements by the method of differential corrections, and an account is given of perturbation techniques. One of these chapters deals with the orbital perturbations of a near-Earth satellite. There is a section on motion in a many-body field and a brief mention of the application of perturbation methods in this case; it is a pity that a description of the solutions which have been obtained in the important case of motion in the Earth-Moon system has been omitted. The closing chapters are briefly concerned with observational techniques—both at optical and radio wavelengths—and with methods of interplanetary guidance and control.

There is an extensive glossary which will be valuable to the reader without a background in astronomy. The non-astronomical definitions should be approached with caution: some, such as 'atomic clock', are not quite precise; while others, such as 'Boltzmann's constant', are incorrect. There is a good bibliography and a total of 70 exercises.

A. P. WILLMORE

GENERAL PHYSICS

Fisica Generale e Sperimentale (*seventh edition, completely revised*), *by E. Perucca. Vol. I, pp. xxviii + 920; vol. II, part I, pp. xii + 618; part II, pp. xi + 619–1228. Unione Tipografico-Editrice, Turin.* 1960. *Vol. I, lire* 12 000; *vol. II* (*both parts*), *lire* 16 500.

For over 25 years this has been the standard Italian treatise of experimental physics, and its ever-youthful author must be warmly congratulated in having produced this new, enlarged, edition, and on having made it completely up to date. It is a vast canvas in the old tradition, outside the present trend of handbooks written by a fairly large number of specialists. While it inevitably lacks the authority of the latter, there are many advantages in a uniform exposition from a single point of view. Moreover the treatment is extremely clear and shows vividly the immense value of a humanistic background in the approach to physics. It is impossible to write a satisfactory short review of a book of this kind. Almost at random, one may remark that it makes a good case for a consistent use of the rationalized MKS system of units; that a judicious balance has been kept in the experimental illustrations of fundamental principles between those of historical interest and their modern counterparts; and that the amount of theory is adequate, given the aim of the book. This aim is to prepare Italian university students for their examination in *Fisica Sperimentale*; it covers broadly the syllabus of the Part I Special Examination in Physics of the University of London. The emphasis, as one would expect, is not always on the same points in the two countries.

The volumes are lavishly produced, the figures being particularly plentiful and useful. L. PINCHERLE

THEORY OF VIBRATIONS

General Dynamics of Vibrations, *by Y. Rocard. Pp. xii + 522. Crosby Lockwood & Sons Ltd, London.* 1960. 80*s. net.*

Until recently almost the only satisfactory treatise on the theory of vibration was that of Professor Den Hartog, of the Massachusetts Institute of Technology. The present book, which is a translation from the third French edition, will undoubtedly become a standard authority on this subject. It covers a very wide field of engineering applications, and discusses them with the aid of adequate mathematical analysis. There are applications to machinery, to the acoustics of the electronics of radio engineering, to gyroscopic theory, to electronic filters, and to the relaxation oscillations of non-linear systems. Even the shimmy of automobile wheels is included. It is regrettable that nearly 17 years have elapsed between the French and English editions, and we must be grateful to W. T. Stern, to whom the first translation of this book is due. G. TEMPLE

ELECTRODYNAMICS

Electrodynamics of Continuous Media, *by L. D. Landau and E. M. Lifshitz (translated from the Russian by J. B. Sykes and J. S. Bell). Pp. x + 417. Pergamon Press, London.* 1960. 84*s. net.*

This volume forms part 8 of a course of theoretical physics, and it is of the high standard that one expects from these two distinguished physicists. It is, indeed, set forth very much in the form of a neat series of lecture notes, and what makes it particularly valuable are the excellent collections of interesting examples, for which full solutions are given.

As it is a book on theoretical physics, experimental details are, of course, omitted. The sections on electrostatics and dielectrics are particularly good, and although some sections—for example, that on the passage of fast particles through matter—are somewhat condensed, physicists will find the volume a most valuable reference book, provided they have an adequate knowledge of tensor analysis. L. F. BATES

CRYSTAL PHYSICS

Imperfections in Crystals, *by H. G. van Bueren. Pp. xviii+676. North-Holland Publishing Co., Amsterdam.* 1960. £5 10*s. net.*

The role that imperfections play in the physical properties of crystals is very important. For some, such as the mechanical strength of metals, it is decisive. The last three decades have seen a systematic study of imperfections leading to the development of a vast field of physics of immense importance to technology. The purpose of this book is to summarize as concisely as possible the theories and the facts.

The first part deals with the properties of lattice imperfections in a general manner, particularly point defects and dislocations. The second part deals in more detail with imperfections in metals, and their relevance to plastic deformation, work-hardening, creep, and so on, and their effect on such important properties as electrical and thermal conductivities. The third part deals with imperfections in non-metallic crystals, particularly the alkali halides, the silver halides, and the industrially important diamond-structure materials, silicon and germanium.

The field is already so vast that much has to be left out if the book is to be made readable. The selection has been skilfully done. The results of theoretical calculations are presented with just enough of the principles of the derivation as is necessary to follow the argument. The diagrams, however, are occasionally so condensed as to be difficult to understand. The experimental evidence is presented succinctly and well illustrated with excellent photographic plates; references to original work are plentiful; and the English is lucid. Graduate students and research workers in solid-state physics will find the book useful, and interested non-specialists will find it readable.

M. H. L. PRYCE

INORGANIC CHEMISTRY

Chemical Periodicity, *by R. T. Sanderson. Pp. x+330. Reinhold Publishing Corporation, New York; Chapman and Hall Ltd, London.* 1960. 94*s. net.*

The rapidly increasing amount of factual information accumulating in inorganic chemistry causes one to look with special interest at any new attempt to classify and correlate these data. In this book the author attempts to make special use of the concept of electronegativity for this purpose. He has published several papers on this subject and has previously drawn attention to many interesting relationships in the Periodic Table.

The book begins with a short discussion of the structure of the atom and of energy levels therein, and then deals briefly with the nature of chemical bonds. This section is followed by a review of electronegativity with a natural emphasis upon the author's own empirical electronegativity values, called stability ratios. He defines the latter as the ratio of the average electron density of the atom to that of a hypothetical iso-electronic inert-gas atom.

In Chapters III and IV topics such as bond polarity, charge distribution in bonds, bond energy, molecular structure, and some aspects of co-ordination chemistry are dealt with.

Then follows a useful chapter on the physical nature of the elements. The remainder of the book is devoted mainly to a discussion of the binary compounds of the elements with oxygen, nitrogen, the halogens, and carbon. The emphasis on these naturally limits the scope of the treatment of other types of compounds.

The chief deficiency of the book is the failure to utilize adequately other complementary principles of physical chemistry, such as the role of lattice energies, hydration energies, and the crystal (ligand) field theory. Also, the relative attention paid to certain topics is puzzling: thus the use of quadrupole coupling constants receives a full page, whereas the important ligand field theory is dismissed in a quarter of a page. There is a tendency to regard complex compounds as a topic rather than as an approach to inorganic chemistry as a whole. Several out of date explanations are encountered; thus, the explanation of square cupric complexes on page 61 is no longer acceptable, and the statement concerning Cr^{+++} complexes on the same page is not correct.

Teachers will find the tabulated data and the graphs very useful, and they will find many of the ideas stimulating and provocative. However, it is another matter whether the author has succeeded in demonstrating for students the fundamental unity of the subject to which he refers in the preface.

R. S. NYHOLM

ORGANOMETALLIC COMPOUNDS

Organo-Metallic Compounds (*second edition*), *by G. E. Coates. Pp. xiii+366. Methuen & Co. Ltd, London; John Wiley & Sons Inc., New York.* 1960. 45*s. net.*

The publication in 1956 of the first edition of this book was warmly welcomed, and since then the volume has been consulted extensively by both undergraduates and research workers. Even in four years, however, so many important developments have taken place that a revision of the text was desirable. The amount of new material has posed a problem. Although the author has been more selective than previously, this second edition is about twice the size of the first, and has outgrown the scope of the series in which it originally appeared. Most of the text has been rewritten. The largest additions concern the sections on the alkali metals, boron, tin, and, in particular, the transition elements. Valency problems, for example those concerning metal-hydrocarbon π complexes, that were mentioned only briefly in the first edition, are discussed at much greater length. In a book of this size the treatment of an expanding subject cannot be completely comprehensive, but the author has succeeded in including an enormous wealth of information. Major developments up to the autumn of 1959 are discussed, and a few references to literature appearing in 1960 have been noted. The author has done a very good job, and this new edition of his book deserves every success.

W. G. OVEREND

ORGANIC CHEMISTRY

Preparativní reakce v organické chemii. V. Aldolisace a příbuzné reakce. (Reactions in Preparative Organic Chemistry. *Vol.* v, Aldol and Similar Reactions), *by J. Plešek and A. Zobáčová.*

Pp. 976. *Czechoslovak Academy of Sciences, Prague.* 1960. *Kčs* 89.

An aldol reaction in the widest sense is any reaction in which a mobile α-hydrogen reacts with a carbonyl or similar group. There are tens of thousands of such reactions, of which 2700 are described in this massive volume, which comprises three main sections. The first is entitled Aldol Reactions and Condensations and deals, *inter alia*, with actual aldol reactions, such as self-additions, intra-molecular condensations, and aldehyde-ketone reactions. The second section is devoted to the Claisen and related types of reactions. The last section deals with Michaels addition reactions and is particularly useful for its chapter on cyanoethylation. All these syntheses are catalysed by bases; but acid-catalysed reactions, such as the Mannich ones, are included at the end of the first section.

2664 references are cited, including patents, and cover the world literature, mainly from 1950 to 1957; there is no bias in favour of the literature of any country. An author and subject index of 35 pages adds to the usefulness of the book. This type of book, where every page is covered with organic formulae, is notoriously difficult to produce, and one must congratulate the editors for having done an excellent job.

M. K. SCHWITZER

STRUCTURAL ORGANIC CHEMISTRY

Optical Rotatory Dispersion, *by C. Djerassi. Pp. xiii*+293. *McGraw-Hill Book Co. Inc., New York; McGraw-Hill Publishing Co. Ltd, London.* 1960. 74*s. net.*

The phenomenon of optical rotatory dispersion—the change of optical rotation with wavelength—has been known for nearly 150 years. The use of this property in structural organic chemistry is, however, very recent. Its application, which has been due to the introduction of new instruments, has been almost entirely the work of one school, that of Professor Carl Djerassi, of Stanford University, California.

This book is the first one on the subject that has been written from the point of view of the organic chemist. It presents a very thorough and yet at the same time a most readable account of the subject, told in a pleasantly personal style. Much of the book is rightly devoted to Djerassi's own work, which has been largely in the field of steroids, terpenes, and related alicyclic compounds. Chapters on other subjects have been prepared by experts in the fields. These chapters are as follows: instrumentation (A. N. James and B. Sjöberg), theory (A. Moscowitz), polypeptides and proteins (E. R. Blout), and amino acids (J. A. Schellman).

Many natural-product chemists will need no introduction to this book. Optical rotatory dispersion is not, however, a specialized tool for some types of natural products; it can be of service in any field of chemistry, organic or inorganic, in which asymmetric compounds occur. Rapid progress is being made at present both in the technique and in its applications.

The production of the book is excellent, and the author, publishers, and printers all deserve congratulation on the speed with which the book was produced.

W. KLYNE

GEOCHEMISTRY

Methods in Geochemistry, *edited by A. A. Smales and L. R. Wager. Pp. vii*+464. *Interscience Publishers Inc., New York; Interscience Publishers Ltd, London.* 1960. $13.50.

This book is an extremely useful guide to the theory, techniques, and applications of the analytical methods used in the collection of geochemical and geochronological data. However, the breadth of treatment implied by the title is not fulfilled, as this is a book on analytical geochemistry only, and does not include other methods, such as those of experimental geochemistry.

The analytical techniques described include refinements of the classical wet chemical methods, colorimetric and flame-photometric methods used in rapid silicate analysis, optical-emission spectrography, fluorescent X-ray spectrography, mass spectrometry (and isotope dilution analysis), radiochemical methods (with a separate chapter on radioactivation analysis), and polarography. An important early chapter describes the techniques of collection and preparation of materials for analysis, and the final chapter is devoted to chemical methods of separation by ion exchange, solvent extraction, and paper chromatography.

The contributions (by twelve authors) are, on the whole, clear and concise. The book is not intended to be comprehensive; for relevant details the reader is constantly referred to the good selection of up to date books, reviews, and papers, given at the end of each chapter.

The book should prove valuable to those who want to know how the analytical data of geochemistry are obtained and the relative values of the methods used in obtaining them.

M. STONE

CELL GROWTH MECHANISMS

Developing Cell Systems and Their Control, *edited by Dorothea Rudnick. Pp. vi*+240. *The Ronald Press Co., New York.* 1960. $8.

The theme of the nine papers presented here is biological differentiation and the control mechanisms that may be involved. Tardent and Bonner deal with whole organisms and discuss regenerating hydroids and developing slime-moulds. Moscona describes some of his own work in one of the more promising fields of modern embryology, the reaggregation of dissociated cells. Hormonal control mechanisms in plants and animals are the subject of two papers by Phinney, West, and Villee, while Singer discusses the curious dependence of amphibian-limb regeneration on the growth of nerves into the limb bud. The last three papers go down to the fundamental mechanisms in cell and molecular biology, whose connections with differentiation are both undoubted and unknown. Work reviews protein synthesis, and Stern describes some interesting experiments on the biochemistry of the cell cycle. Novikoff's paper is a lively and stimulating account of his cytochemical work and is an example of the good that can be done in a field that is often pedestrian. But it has little direct connection with the main theme of the book, and this emphasizes the only major criticism that can be made about this symposium. Differentiation is a vast, amorphous, and very difficult subject, and it would have been better, within a limited framework of nine papers, to concentrate on one level of organization, rather than to try to bridge the gap between animals and molecules.

J. M. MITCHISON

ENZYMOLOGY

Advances in Enzymology, Vol. XXII, *edited by F. F. Nord. Pp. v*+567. *Interscience Publishers Inc., New York; Interscience Publishers Ltd, London.* 1960. £5 6*s. net.*

All enzymologists and most other biochemists look forward to the annual appearance of 'Advances in Enzymology'.

The latest volume presents a rich diet that should disappoint nobody.

The contents of volume XXII cover a wide range, from pure enzymology (the active site and enzyme action, synthesis and hydrolysis of sulphate esters, coenzyme binding, and enzyme chromatography) to chemical metabolism and genetics (the biosynthesis of cholesterol, the biochemistry of sulphonium compounds, the induced synthesis of proteins, and genetically controlled differences in enzyme activity). We also find a most interesting article on the synthesis of nucleotide coenzymes, written by two organic chemists.

There is no question in the reviewer's mind as to the high quality of all the articles, but it does seem odd that some of the topics are included in a specialized volume on enzymology. However, the subtitle is 'and related subjects in biochemistry'. The authors write for the general biochemist and yet succeed in being helpful to the specialist. Obviously, the balance between these two aims differs somewhat from chapter to chapter. H. GUTFREUND

COMPARATIVE ANATOMY AND TAXONOMY

Primates, Comparative Anatomy and Taxonomy, *Vol. IV*, Cebidae, *Part A, by W. C. Osman Hill. Pp.* xxii + 523. *Edinburgh University Press, Edinburgh.* 1960. £9 9*s. net.*

Dr. Osman Hill's encyclopaedic work on the comparative anatomy and taxonomy of the Primates has now reached its fourth volume with the publication of Part A of his monograph on the Cebidae. This deals with four of the sub-families of the New World monkeys. Like the previous volumes, it is mainly concerned with the analysis of the external and internal anatomical characters of the various genera and species, but it also includes accounts of habitat, diet, and behaviour. One can hardly fail to be impressed with the assiduity of the author and with the immense labour that he has undertaken to sift the voluminous literature for all the information he presents in this compendious work. More than this, he displays an intimate personal knowledge of his material, and his volumes include many original new observations of importance. To have embarked on so comprehensive a series of monographs on the Primates implies considerable courage as well as sustained effort, and the results that have so far been published are to be highly commended. The text is accompanied by a profusion of illustrations of high technical quality, and the style of writing is precise and to the point. A standard work of reference of this kind, which aims at being all-inclusive, means that in some instances data are recorded on the basis of single specimens, so that it has not been possible to indicate the range of variation in the anatomical characters of each individual species; this cannot reasonably be expected. In the course of time we look forward to the completion of this encyclopaedia with successive volumes on the other subfamilies of the Cebidae, and on the Old World monkeys and anthropoid apes. Together, they will certainly comprise the most up to date work of reference on the Primates, and will be found essential not only for primatologists, but for zoologists generally.

SIR WILFRID LE GROS CLARK

HISTORY OF TECHNOLOGY

Transactions of the Newcomen Society for the Study of the History of Engineering and Technology, *Vol. XXX* (1955–56 and 1956–57). *Pp.* 281 + 62 *plates. The Newcomen Society, London.* 1960. 50*s. net.*

Like its predecessors, this volume of the Newcomen *Transactions* contains such a wealth and variety of papers on the history of technology that one may comment here only on those of the widest current interest.

The recent studies of Professor Derek J. de Solla Price on the Antikythera mechanism make Dr Von Bertele's paper on the 'Origin of the Differential Gear and its Connection with Equation Clocks' of special interest, for it clears up much of the confusion between a mechanical astronomical computer as part of a clock and the actual mechanical time-measuring device itself, at least in the gearing required. Is not the time now ripe for someone to give us a first-rate general history of gears and gearing?

We are especially indebted to Mr A. Stowers, whose presidential address gives us drawings and dimensions on the basis of which a more tightly reasoned analysis of the Roman artifacts at Chollerford Bridge, Chesters, Northumberland, may be made. Professor Richmond's analysis, referred to in Vol. XXIX, is less than satisfying to any engineer familiar with the actual objects. The plan of the mill shows that only a water wheel with a vertical axle can fit the data, and the known fact that the Roman garrison in this area came from Syria accounts for this early appearance of the Greek mill in Britain. One need also only set the 'hub' with its axis vertical, note the lack of any kind of bearings, and examine the carefully dressed end surfaces to recognize it as a barrel hammer, worked by a group of men with staves, intended to produce the firm masonry characteristic of the bridge abutment.

It is a pity that the long delay in appearance of these papers in print has prevented earlier and more general discussion of their excellent material.

R. S. WOODBURY

Pilkington Brothers and the Glass Industry, *by T. C. Barker. Pp.* 296. *George Allen and Unwin Ltd, London.* 1960. 40*s. net.*

An outstanding firm has found the historian it deserves. He has traced the vicissitudes through which the operation of a single 120 ft cone, built in 1826 to make crown glass, has grown into a concern that controls plate- and sheet-glass manufacture throughout Britain. Pilkingtons is still a private limited company, but family piety has not been allowed to obscure the picture of Victorian industrial relations. Thus, after an unsuccessful strike, the new wages were to be 'the lowest than can possibly be gone to', with staggered contracts for skilled workers and a 10s. fine for a day's absence without cause. Even so, the lower wages and lower standard of living on the continent were among the factors that made the Belgians their most dangerous trade rivals.

In each generation the firm has faced a new crisis of technological advance. The latest of these, the establishment of their float process for continuous production of flat glass, involved 'the chemical and physical equilibrium problems between hot liquid glass and hot liquid metal and hot atmosphere'. It took eight years, three new plants, and £4 000 000 to find the perfect solution.

T. K. DERRY

Short notices of books

(These notices are descriptive rather than critical and are designed to give a general indication of the nature and scope of the books.)

A Dictionary of Named Effects and Laws in Chemistry, Physics and Mathematics (*Second edition*), *by D. W. G. Ballentyne and L. E. Q. Walker. Pp. v + 234. Chapman and Hall Ltd, London.* 1961. 30*s. net.*

This book is concerned with effects and laws that have been named after those who discovered or studied them. The second edition contains more than 150 new definitions, and the appendix on units has been rewritten.

Physics and Archaeology, *by M. J. Aitken. Pp. x + 181. Interscience Publishers Inc., New York; Interscience Publishers Ltd, London.* 1961. 43*s. net.*

The author of this book has aimed at describing some of the ways in which physics has been applied to archaeology. It is intended for both the general scientific reader and the student of archeometry (measurement of archaeological materials), and the chapters are arranged so that one technique, and even one application of a technique, can be studied more or less without reference to others. The techniques described include magnetic location, resistivity surveying, and radiocarbon and magnetic dating.

Russian Journal of Physical Chemistry, *No.* 7, 1959, *edited by Ya. I. Gerasimov. Pp.* 107. *The Chemical Society, London.* £30 *per annum.*

This is the first issue of an English translation of Журнал физической Химии. The translation will be published monthly, under the editorship of R. P. Bell, and will appear approximately three months after the Russian issue. Papers in this issue include 'Thermodynamic properties of sodium compounds in the solid state', by P. G. Maslov; 'Mutual influence of substituents in *p*-disubstituted benzene derivatives', by P. P. Shorygin; and 'Kinetics of thermal dissociation of diatomic molecules', by E. V. Stupochenko and A. I. Osipov.

Small Scale Experimental Chemistry, *by T. A. H. Peacocke. Pp. xii + 164. Longmans, Green & Co. Ltd, London.* 1960. 8*s.* 6*d. net.*

This book aims at providing a course of experiments to teach—using the heuristic method—the principles of qualitative inorganic analysis, and a series of small-scale inorganic preparations, such as are required for the General Certificate of Education in Britain. A certain amount of traditional usage has been discarded. Thus equations are, in general, written in ionic form, and the modern versions of the formulae of such oxyanions as zincate and aluminate are used. The emphasis is on experiments on a semimicro scale, and an appendix describes methods of making some of the apparatus that is needed.

An Introduction to Transition-Metal Chemistry; Ligand-Field Theory, *by L. E. Orgel. Pp.* 180. *Methuen & Co. Ltd, London; John Wiley & Sons Inc., New York.* 1960. 25*s. net.*

The intention of the author of this book is that it will be used by students of inorganic chemistry and others seeking a general survey of ligand-field theory, and the book is therefore both short and non-mathematical. The various chapters describe, among other subjects, orbitals and energies, stereochemistry, reaction mechanisms and reaction rates, and the complexes formed by unsaturated hydrocarbons.

Advanced Organic Chemistry (*third edition*), *by G. W. Wheland. Pp. xi + 871. John Wiley & Sons Inc., New York; John Wiley & Sons Ltd, London.* 1960. £7 *net.*

This textbook aims at giving, in considerable detail, structural theory in its broadest sense; it deliberately omits discussion of descriptive organic chemistry. The author has continued to emphasize resonance theory in this edition, and he has added such topics as electronic and nuclear magnetic resonance, conformational analysis, and the Hammett rho-sigma relations. He has also extended the discussion of reaction mechanisms and indicated the value of kinetic data to his approach.

Name Index of Organic Reactions (*second edition*), *by J. E. Gowan and T. S. Wheeler. Pp. iii + 293. Longmans, Green & Co. Ltd, London.* 1960. 50*s. net.*

A large number of organic chemical reactions are referred to by the name of their discoverer, and this book has been produced with a view to combating the obscurity that this practice can produce. The second edition is considerably larger than the first, and aims at being comprehensive; it also now includes formulae. An index of reaction types is included.

The Chemistry of the Terpenes, *by A. R. Pinder. Pp. vii + 223. Chapman and Hall Ltd, London.* 1960. 50*s. net.*

The aim of this book is to be a bridge between general textbooks of chemistry and comprehensive books on terpenes. Particular stress has been laid on the reactions involved in degradation and synthesis, the chemical transformations that occur in terpene chemistry, and some recent developments in stereochemistry and biosynthesis. The subjects covered include the various monoterpenes, diterpenes, triterpenes, tetraterpenes, rubber, and the biogenesis of terpenes.

Aerosols: Science and Technology, *edited by H. R. Shepherd. Pp. xiv + 548. Interscience Publishers Inc., New York; Interscience Publishers Ltd, London.* 1961. £8 9*s. net.*

Essentially, this book is a survey of present practice in aerosol technology. There are chapters dealing with the theory and practice of aerosols, and with the containers, valves, and propellents used. Other chapters are concerned with materials that are prepared nowadays in aerosol form, such as cosmetics, pharmaceutical products, foods, pesticides, and paints.

Biographical Memoirs of Fellows of The Royal Society, *Vol. VI. Pp.* 295. *The Royal Society, London.* 1960. 30*s. net.*

These memoirs are written by various authors and aim at giving a biography of the subject, together with an assessment of his work. The current volume includes memoirs of A. C. G. Egerton by D. M. Newitt, I. M. Heilbron by A. H. Cook, Max von Laue by P. P. Ewald, and F. A. Paneth by H. J. Emeléus.